Cooperation and social control

Bakker, Dieko Marnix

DOI:
10.33612/diss.98552819

2

A comparison of three measures of Social Value Orientation

Dieko M. Bakker, Jacob Dijkstra

This chapter is currently under review at an international peer-reviewed journal

Social Value Orientation (SVO) is one of the most frequently studied individual traits in research on social dilemmas (Au & Kwong, 2004) and one of the most vital to understand and measure for research in this field (Murphy & Ackermann, 2012; Murphy et al., 2011). SVO literature (e.g. Au & Kwong, 2004; Balliet et al., 2009; Bogaert et al., 2008; Van Lange et al., 2013) shows that SVO consistently relates to behavior in social dilemmas. In experimental studies on common-pool resources, participants with different Social Value Orientations take different amounts from the common pool (Au & Kwong, 2004; Liebrand, 1984). This finding is robust to changes in the structure of the common-pool-resource game (Au & Kwong, 2004). SVO also correlates with contributions in public good games (e.g. De Cremer & Van Lange, 2001; Dijkstra & Bakker, 2017; Fung, Au, Hu, & Shi, 2012), Prisoner's Dilemma games (e.g. Murphy, Ackermann, & Handgraaf, 2011), investment games (e.g. Kanagaretnam, Mestelman, Nainar, & Shehata, 2009) and various other types of social dilemma games (Balliet et al., 2009). One part of these differences in behavior may be a direct consequence of differences in SVO, and another part may be due to different expectations regarding the behavior of others between persons with different SVO types (Pletzer et al., 2018). Additionally, there is evidence from non-experimental studies suggesting that a person's Social Value Orientation influences, for example, volunteering behavior (Au & Kwong, 2004; Pletzer et al., 2018), donations to noble causes (Pletzer et al., 2018; Van Lange et al., 2007) and pro-environmental behavior (Pletzer et al., 2018; Van Vugt et al., 1996). SVO relates not just to behavior but also to participation in experiments. Prosocials are more likely to volunteer for experiments than individualists and competitors (Van Lange, Schippers, & Balliet, 2011). Additionally, the distribution of Social Value Orientations in experimental samples may depend on properties of the groups (of e.g. students) from which these samples are drawn. Van Lange et al. (2011) showed that while the prosocial orientation is the most common orientation among psychology students and in representative samples of the Dutch population, individualists are the largest group among economics students.

An accurate measurement of Social Value Orientation is thus crucial for our understanding of behavior in social dilemmas. Several measures of SVO are available (Au & Kwong, 2004; Murphy & Ackermann, 2012) but there are few comparative studies. As a result, we do not have a clear picture of how different measures of Social Value Orientation relate to each other, nor do we know definitively whether some measures are better than others. Increasing our knowledge of the relationships between the three most prominent measures and helping us choose between them are the main aims of this study.

Defining Social Value Orientations

Central to the concept of Social Value Orientation is the observation that the behavior many people exhibit in social dilemmas is not solely aimed at the maximization of their own material gain. Instead, a significant proportion of people in such situations show consideration for the welfare of others (Au & Kwong, 2004; Balliet et al., 2009; Bogaert et al., 2008; Van Lange et al., 2013).

SVO is defined in terms of weights individuals assign to their own and other's outcomes in situations of interdependence (Messick and McClintock 1968). Most commonly, three or four types of SVO are distinguished (Au & Kwong, 2004; Bogaert et al., 2008; Murphy & Ackermann, 2012). The first of these is the altruistic orientation. Altruists care positively about the outcomes of others and are neutral about their own outcomes. More intuitively, altruists try to reach the most positive outcome possible for the other without regard to the outcome for themselves. Cooperators, the second commonly distinguished group, care both about their own outcomes and the outcomes of the other. They typically try either to maximize the joint outcomes for themselves and the other or to minimize the difference in outcomes between themselves and the other (Van Lange et al., 2013). Altruists and cooperators are sometimes grouped together as one prosocial orientation (Bogaert et al., 2008; Murphy & Ackermann, 2012). All individuals who care positively about (i.e. place a positive weight on) the outcomes of others can be considered prosocial. The third commonly distinguished category is the individualistic orientation. Individualists care about obtaining the most advantageous outcome for themselves (i.e. they assign a positive weight to their own outcome) regardless of the outcome for the other. Competitive individuals, finally, care positively about their own outcome and negatively about the outcome for the other. That is, they try to obtain the maximum relative advantage possible compared to the other. The various orientations are characterized by the weights individuals place on the outcomes of self and other, as well as by their inferred motivations and typical payoff allocations (Murphy & Ackermann, 2012, Table 2). Several other, less common, orientations have also been identified (Au & Kwong, 2004; Murphy & Ackermann, 2012). However, many studies incorporating measures of Social Value Orientation do not make use of these orientations and the vast majority of individuals can be classified into the four most common categories.

Traditionally, Social Value Orientation has been used as a categorical construct, classifying respondents into one of the SVO orientations and using the respondent's orientation as a predictor of behavior (Murphy & Ackermann, 2012). However, there

may be distinct subcategories within an orientation, such as differences among prosocials in whether they are mainly concerned with the maximization of joint outcomes or with the minimization of (advantageous) inequality (Murphy et al., 2011; Van Lange et al., 2013). Additionally, the variation between respondents with the same classification which can be observed when SVO is measured on a continuous scale suggests that many individuals are not purely prosocial, individualistic or competitive. Rather, there are more gradual differences which are also accompanied by gradual differences in outcomes typically associated with SVO (Murphy & Ackermann, 2012).

Aims of the paper

Because SVO is very frequently used as a predictor of behavior, and because there are different types of measures which are supposed to measure the same concept but may be unintentionally different, systematically comparing different measures of SVO is important. As identified in a recent meta-analysis by Pletzer et al. (2018), the three most commonly used measures of SVO are the 9-item Triple Dominance Measure (TDM; Van Lange et al., 1997), the Ring Measure (RM; Liebrand & McClintock, 1988) and the Slider Measure (SLM; Murphy et al., 2011). The Slider Measure is the most recent of the three and appears to be replacing the Triple Dominance Measure and the Ring Measure. Reviews published before the introduction of the Slider Measure recognize the Triple Dominance Measure and the Ring Measure as the most common ways to measure SVO (Au & Kwong, 2004; Bogaert et al., 2008).

An excellent overview of the benefits and drawbacks of many types of SVO measures has been provided by Murphy & Ackermann (2012). This overview discusses the validity and reliability of these measures, as well as their output resolution (the ability to distinguish nuances in SVO), efficiency (in terms of time and effort to both complete and evaluate the measure), and unique features (Murphy & Ackermann, 2012). As Murphy & Ackermann (2012) note, however, there are few studies which perform systematic empirical evaluations of and comparisons between measures of SVO. The Slider Measure, in particular, has only been evaluated by its original authors. Independent replication of its good psychometric properties is valuable. Additionally, there are several properties of the three measures which we believe have not been systematically investigated before. In this study we thus address four topics, which in our opinion have not yet been sufficiently addressed in the SVO literature.

First, we investigate the sensitivity of the three measures to random responses. While previous reviews have discussed the exclusion of invalid responses when

describing how cases are classified by each measure (e.g. Au & Kwong, 2004; Murphy & Ackermann, 2012), we are not aware of any study which has theoretically and empirically assessed how successfully each measure discriminates between random and genuine responses. Outcomes on each measure of Social Value Orientation may be influenced not only by a person's orientation but also by properties of the measure itself. One way to investigate differences in the properties of the three measures is to investigate how they classify completely random responses. There are two ways in which this can reveal systematic differences between the three measures. For one, measures of SVO often try to distinguish between valid (representing a person's true SVO) and invalid (random or otherwise non-serious) responses. By investigating the classification of random responses we can compare the effectiveness of exclusion criteria between the three measures. For another, looking at the distribution of classified random responses reveals differences between the measures in the probability that a random response is classified as altruistic, cooperative, individualistic or sadistic. This indicates certain "tendencies" a measure has of classifying a response in either one category or another. Overall, good measures should be able to distinguish genuine responses from random ones without leaving a large proportion of respondents unclassified, and should not "steer" responses in a "preferred direction".

Second, we investigate the convergent validity of the three measures. Little evidence is available on whether the different measures of SVO assign the same classification to the same respondent (Au & Kwong, 2004; Bogaert et al., 2008). We know of only one previous study which has compared all three measures within the same sample (Murphy et al., 2011) and in that case, each respondent only completed two of the three measures.

Third, we investigate the test-retest reliability of the three measures over a period of approximately three months. Again we know of only one previous study to evaluate the test-retest reliability of all three measures (Murphy et al., 2011) and in that study, the three measures were evaluated over much shorter and, moreover, unequal intervals than in our study.

Fourth, we will pay particular attention to the choice between categorical and continuous measures of Social Value Orientation. Recent reviews of the literature suggest moving from categorical to continuous measures (Bogaert et al., 2008; Murphy & Ackermann, 2012; Murphy et al., 2011; Pletzer et al., 2018), for both theoretical and empirical reasons. On the theoretical side of the argument, Social Value Orientation can be considered a continuous construct given that it is defined in terms of the relative importance individuals attach to the payoffs of others and to

their own payoffs (Murphy et al., 2011). In principle, these relative weights could take any value and there is no obvious reason to restrict them to predefined ideal types. On the empirical side, when Social Value Orientation is measured on a continuous scale it appears that there is substantial variation in responses which would be discarded if the measure were reduced to categories (Murphy et al., 2011).

We will first go into more detail on the measures of Social Value Orientation we will evaluate. We will then evaluate the three most common measures of SVO, including a more comprehensive overview of the literature on each of the topics we address, and conclude with our recommendations for future studies.

MEASURING SOCIAL VALUE ORIENTATION

Measures of Social Value Orientation ask respondents to choose between several alternative allocations of money, points, or resources between themselves and an anonymous other. The respondent's chosen allocations are used to estimate the weights they attach to their own outcomes and the outcomes for the other. Measures usually include several similar items, intended to more clearly distinguish between persons with different orientations and (in the case of measures which can be used as continuous outcomes) make a more accurate determination of a person's exact orientation. We will explain the concept and procedures of each of the measures we include in this study. The examples presented in this section are also used in the questionnaires completed by our respondents.[1]

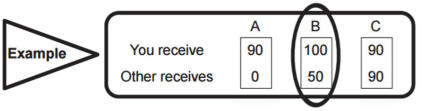
9-Item Triple Dominance measure

Practically, the 9-item TDM (Van Lange et al., 1997) is the simplest of the commonly used measures. It has just nine items, each with only three alternatives to choose from, where each alternative clearly represents one of the three SVO orientations. The classification rule, which states that at least 6 out of the 9 items must be answered consistently for a participant to be classified (Van Lange et al., 1997), largely prevents participants who employ random answer patterns from being treated as if they

[1] These questionnaires (paper-based examples of the TDM, RM and SLM) used to be available at http://vlab.ethz.ch/svo/SVO_Slider/SVO_Slider_paper_based_measures.html and were downloaded from this website in 2015. This URL is now defunct. Examples and translations for the paper-based Slider Measure can still be downloaded at http://ryanomurphy.com/styled-2/downloads/index.html. The Slider Measure we used corresponds to Version A on that page, with adapted instructions. The measures and instructions we used are available at https://osf.io/6rdx9/?view_only=87831a672837458eb667abe89bc818e1.

legitimately indicated their SVO. On the other hand, this classification rule often leaves a substantial number of people unclassified (Au & Kwong, 2004) and does not allow for the presence of mixed SVO types. Figure 2.1 shows an example choice from the 9-item TDM.

Figure 2.1. Example from the 9-item Triple Dominance Measure



The 9-item TDM is designed to be used as a purely categorical measure. Researchers who have recognized the benefit of a continuous measure of SVO have tried to extract continuous information from the 9-item TDM, but these transformations are discouraged because they risk confusing the reliability of a preference with its magnitude (Murphy & Ackermann, 2012).

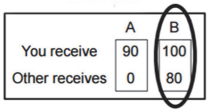
Ring measure

The ring measure (RM) of SVO is presented to participants as a set of items with payoff pairs between which they are expected to choose (the set contains 24 items, each with two payoff pairs) (Au & Kwong, 2004). A payoff pair is an ordered pair of payoffs for self and other. The payoff pairs presented to participants are derived from equally-spaced points on a circle with a fixed radius, whereby one axis (usually the horizontal axis) represents payoffs for self and the other axis (usually the vertical axis) represents payoffs for other. A person's SVO orientation is also a point on this circle, which represents the person's ideal payoff combination. The idea behind the RM is that a person will choose the own-other payoff combination closest to their ideal payoff combination, which represents their SVO (Au & Kwong, 2004; Liebrand & McClintock, 1988). Based on the total allocation to self and the total allocation to other, the person's SVO can be computed as a point defined by an angle (representing the relative weight of payoffs to self and to other) and a vector length (representing how consistently responses indicate a single SVO) from the origin of the circle.

This angle can be used as a continuous measure of SVO when only the right half of the ring, with positive payoffs for self, is used (Murphy & Ackermann, 2012). Traditionally, however, responses on the RM are reduced to categorical classifications. Liebrand (1984) suggested dividing the circle into eight equal octants denoting eight

social value orientations (see also Table 2 in Au & Kwong, 2004 or Figure 1 in Murphy & Ackermann, 2012). In recent studies, the vector length is used as a consistency indicator, with vectors shorter than a quarter of the maximum length possible (i.e. twice the radius of the ring) considered inconsistent (Au & Kwong, 2004). Earlier studies tended to use 50 or 60 percent agreement (i.e. responses indicating the same SVO orientation) as a cutoff for consistency. In seventeen studies identified by Au & Kwong (2004), ten used this 50 or 60 percent criterion and the rest used a minimum vector length of 25 or 20 percent of the maximum. Figure 2.2 shows an example choice from the ring measure.

Figure 2.2. Example from the ring measure



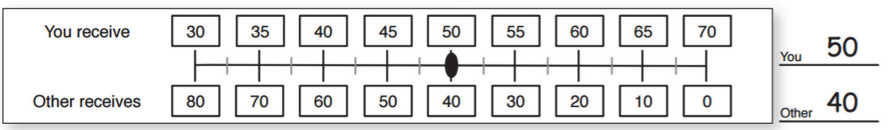
Several variations of the Ring Measure exist. The questionnaire we used is a half ring (Murphy & Ackermann, 2012), which only allows Altruistic, Cooperative, Individualistic, Competitive and Sadistic orientations (the right half of the ring, including the most common orientations). A complete ring would also allow for orientations in which the respondent places a negative weight on their own outcome, meaning that the respondent is to some extent masochistic (Au & Kwong, 2004; Murphy & Ackermann, 2012). However, such responses are exceedingly rare and these categories are not usually reported (Au & Kwong, 2004; Bogaert et al., 2008). There are several advantages to using only the most common (right) half of the ring. For one, including choices which are intended to distinguish between very uncommon orientations is inefficient and may result in inconsistent choices (Murphy & Ackermann, 2012). For another, using only the right half of the ring allows an interpretation of the SVO angle as the weight an individual attaches to the other's outcome relative to their own (Murphy & Ackermann, 2012).

Slider measure

The slider measure (SLM) of social value orientation is a more recent measure, proposed by Murphy et al. (2011). The authors claim that existing methods are inefficient (e.g. including items with very little variation in choices) and often fail to produce consistent results for a substantial proportion of subjects or require substantial time and effort on the part of participants. Additionally, they claim that

existing measures have not been explicitly designed to capture more nuanced motives such as inequality aversion (Murphy & Ackermann, 2012; Murphy et al., 2011). The authors state that SVO should be assessed on a continuous scale because SVO is a continuous construct, which represents how individuals balance their own outcomes and those of others. The existing measures of SVO which produce mainly categorical data (e.g. the 9-item TDM) are therefore missing a substantial amount of information regarding peoples' social preferences (Murphy et al., 2011). For this reason, the authors have designed the slider measure to enable measurement of SVO on a continuous scale. The slider measure asks participants to choose a resource allocation over a continuum of joint payoffs. The slider measure consists of six primary decisions measuring basic SVO, with the addition of nine secondary decisions. The secondary items can be used to disentangle inequality aversion (minimizing payoff differences between self and other) and joint gain maximization (maximizing the sum of payoffs to self and other). Compared to other measures of SVO, the authors claim that the SLM allows researchers to 1) evaluate whether respondents understood the task, 2) evaluate the transitivity of preferences as an indicator of genuine responses, 3) create a complete rank order of all orientations for each respondent and 4) score the measure in such a way as to yield a single continuous index of SVO (Murphy et al., 2011). The index produced by this measure can be coerced to categorical values, as with other SVO measures, but can also be used in its continuous form.

Figure 2.3. Example from the slider measure



The slider measure itself can be used with a continuous choice scale (most suitable when respondents participate using a computer or similar means) or with a set of nine discrete choices (most suitable when respondents participate using a pen-and-paper questionnaire). In either case, participants are classified through a procedure similar to the ring measure. First, the mean allocations to self and other are calculated. Then, these means are adjusted so that the computed SVO angle will originate from the center of the circle described by the Slider Measure. The ratio between the adjusted mean allocation to other and the adjusted mean allocation to self describes the

tangent of the SVO angle, so the angle is computed as the inverse of this tangent (Murphy et al., 2011).

To assess the consistency of responses to the Slider Measure, the authors recommend checking the transitivity of respondents' preferences (Murphy & Ackermann, 2012; Murphy et al., 2011). Transitivity of preferences entails that when a person prefers orientation A over B and prefers B over C, this person also prefers A over C. Genuine responses to the SLM are supposed to produce transitive social preference choices.

DATA

Variables

We asked our respondents to complete each of the three SVO measures. For the Slider measure, we asked respondents to fill out both the six primary items to determine their SVO and the 9 additional items to distinguish between prosocial motives. Additionally, we asked each respondent to indicate their age, their gender (male/female/other), their year of studies (first year or second year), their most recent prior education/occupation and their currently obtained number of course credits (ECTS).

Data collection

Data were collected among sociology students at the University of Groningen, The Netherlands. The department of sociology at this university requires students in the first and second year of their bachelor's to participate in sociological research at so-called 'Test Days'. These days offer researchers the opportunity to gather data while familiarizing students with the practice of sociological research. These Test Days take place twice a year, once in the fall and once in the spring. Since students are required to participate in their first and second year, this means that each student takes part in four consecutive Test Days. We included our survey in four consecutive test days in order to investigate the consistency of SVO classifications over time within a stable and relatively homogeneous sample. The first wave of data collection took place from the 12th of November 2015 to the 21st of November 2015. The second wave took place from the 22nd of February 2016 to the 14th of March 2016 (roughly three months after the first wave). The third wave of data collection took place from the 21st of October 2016 to the 26th of October 2016. The fourth wave took place from the 10th of March 2017 to the 20th of March 2017.

At each Test Day, we distributed questionnaires which included the three measures of social value orientation as well as several control variables regarding the personal characteristics of the student. Each questionnaire started with these control questions, followed by the SVO measures. Six versions of the questionnaire were distributed, each with a different order of SVO measures, so that all possible orders were represented. Students were not assigned an order in advance. Which version of the questionnaire they answered was determined by the desk at which they happened to sit down. All versions of the questionnaire as well as the codebook are hosted on the Open Science Foundation framework (see footnote 1). Students are not paid for participating in these Test Days. Our SVO measures were therefore not incentivized.

Students were always allowed to decline to answer some or any questions. For each control question, a 'No answer' option was provided. For the SVO measures, we provided one 'No answer to any question in this section' checkbox at the start of each measure. Additionally, students were asked to sign a release form at the end of their Test Day session, which entitles the researchers to use the student's answers. The responses of any students who did not sign this release form or indicated a desire that their answers not be used were not presented to the researchers and are therefore not included. Prior to the first wave of data collection, this study was approved by the ethics committee of the University of Groningen's sociology department, as is required for all studies conducted at the Test Days.

Sample

A total of 110 respondents participated in the first wave of this study. Of these 110 respondents, 44 were male and 66 were female (an option for other genders was provided but not selected by any respondents). The sample consisted of 52 first-year students and 58 second year students. The average age of these respondents was 19.83 years (*SD* = 1.64).

In the second wave, 97 respondents participated. Of these, 89 had also participated in the first wave. The sample composition in the second wave was somewhat different with 33 male versus 64 female students and 38 first-year students versus 59 second year students. The sample suffered from attrition mainly among male first-year students. This group is almost entirely responsible for the decreased number of respondents in the second wave. We know from student records that this cohort of first-year students had an unusually high dropout rate, apparently mainly among male students.

When we include also the third and fourth wave, we have a subset of 22 students whose Social Value Orientation has been measured at all four time-points. The sample of students who participated in both the first and the fourth wave is slightly larger, with a total of 27 respondents. Because the subset of participants who participated in all four waves (or at least in both the first and the last wave) is small, we mostly omit discussion of long-term comparisons. For our analysis of changes over time, we focus on comparing the first and second wave, which are approximately three months apart. This time interval is much larger than intervals in other comparative studies. Descriptive analyses are available in an online appendix (see footnote 1).

Classification

The three measures we assess in this study (SLM, RM, and 9-item TDM) each have a different method of classifying participants according to their SVO type. Classification is simplest in the 9-item TDM. Each of the nine items of this measure has a prosocial option, an individualistic option, and a competitive option. Participants are classified as a certain type when they choose the option corresponding to that type on at least six out of the nine items. The SLM and RM both use the respondent's chosen allocations to compute an angle which represents the respondent's SVO. This angle indicates the respondent's ideal balance between their own payoff and the other's payoff. Both measures offer cutoff values which can be used to transform these angles into the commonly used categorical classifications. The cutoff values for each measure are presented in Table 2.1.

Table 2.1. Cutoff values

	Slider		Ring	
	From	*To*	*From*	*To*
Altruistic	57.15°	-	67.50°	112.50°
Cooperative	22.45°	57.15°	22.50°	67.50°
Individualistic	-12.04°	22.45°	-22.50°	22.50°
Competitive	-	-12.04°	-67.50°	-22.50°
Sadistic			-112.50°	-67.50°

In our analysis of the data, we will often use these categorical classifications, particularly when assessing to what extent the three measures result in similar classifications. However, we will also devote some time to exploring the potential downsides of transforming the SVO angle to a categorical classification.

Within the prosocial type, it is common to distinguish between a cooperative type (those who maximize joint outcomes or minimize inequality) and an altruistic type (those who maximize the other's payoff) (Au & Kwong, 2004). The 9-item Triple Dominance Measure does not make this distinction. The Slider Measure does make this distinction, but calls these categories 'Altruists' and 'Prosocials' (Murphy et al., 2011). In this paper we will use the terms 'altruists/altruistic' and 'cooperators/cooperative' to refer to the subtypes, and use the term 'prosocials/prosocial' to refer to the combined type.

Analysis plan

We present our analyses in two parts. First, we look at how the three measures classify random answering patterns to assess how well each measure manages to distinguish between genuine and random responses. Then, we use the first unique observation from each respondent to empirically assess the convergent validity of the three measures. Finally, we assess the test-retest reliability of classifications from the first to the second wave (a period of three months between tests). We investigate, for each measure, how similarly respondents score on each of the three measures in the first and second wave. For this, we use all respondents who took part in both the first and the second wave.

ROBUSTNESS OF CLASSIFICATIONS TO RESPONSES

Measures of Social Value Orientation commonly include a consistency criterion intended to exclude invalid or unclassifiable responses. These criteria should prevent respondents whose answers are not clearly consistent with one of the orientations from being classified. Such criteria have to balance false positives and false negatives. On the one hand, when too stringent a criterion is applied we may exclude more valid responses than necessary. On the other hand, too loose a criterion will allow invalid responses into our samples and negatively impact the quality of our data. Respondents who do not take their participation seriously and respond at random are a likely source of false positives and a useful benchmark. By investigating the performance of each measure against random responses we can gain some insight into the effectiveness of their exclusion criteria.

There are two issues when assessing the balance between false positives and false negatives in each measure. First, responding randomly is only one of many ways to respond to measures of Social Value Orientation without answering in accordance with one's true SVO. In fact, it may be more likely that such responses are not truly

random but rather follow some predictable pattern (e.g. always selecting the first option). Because each of the three measures can be used in different variations, with for example different orders for both items and responses, it is difficult to reasonably judge how predictable patterns can affect responses in general. Performance against truly random responses, on the other hand, does not depend on the specific order of the items and response options of each measure. For this reason, we now focus on the classification of random responses, and investigate performance against predictable answering patterns for the specific implementations of each measure that we used in an online appendix (see footnote 1).

Second, while assessing the probability of false positives in this way is relatively straightforward, estimating the probability of false negatives is a more difficult task. We would need a response model for genuine responses which includes the possibility that even respondents who attempt to answer in accordance with their true SVO sometimes make a mistake. With the available information and assuming any errors real respondents make are random, we can place an upper limit on the false negative rate in our sample, which is equal to the percentage of cases which remain unclassified.

Method

For measures which use discrete choices, we are able to enumerate all possible combinations of choices and assess how people would be classified based on each of these decision profiles. This applies to the 9-item TDM, the RM and the discrete version of the SLM. For the 9-item Triple Dominance Measure, we can also calculate directly how likely it is that a participant who gives random answers is classified as each type, or remains unclassified. For measures which use a continuous scale, we cannot enumerate all possible decision profiles. This applies to the continuous version of the Slider Measure. For the continuous SLM, we simulate a large sample of possible combinations of allocations to cover the decision space. The results of this simulation are omitted since they differ little from those obtained for the discrete SLM. Scripts for the enumeration and classification of decision profiles are available on the Open Science Framework (see footnote 1). All scripts are written in R (R Core Team, 2017).

9-item Triple Dominance Measure

Because the 9-item Triple Dominance Measure uses a simple classification system whereby each choice clearly represents a certain type, and a person is classified by consistently (6 or more times out of 9) selecting a given type, we can calculate directly

how likely each classification is for a respondent who gives random answers. The probability of picking a particular type on a particular item is simply equal to 1/3. The probability of being classified as a particular type, $P(C_t)$, with the specified classification threshold of 6 consistent answers is:

$$P(C_t) = P(X \geq 6) = \sum_{i=6}^{9} \binom{9}{i} \frac{1^i}{3} \times \frac{2^{9-i}}{3}$$

Which when evaluated results in a probability of 0.04242 of being classified as prosocial, the same probability for individualistic and competitive, and a residual probability of 0.8727 of being considered unclassified. If we vary the classification threshold from 5 (the lowest possible threshold for which a respondent cannot be classified as two types at once) to 9, Table 2.2 shows us how the probabilities of classification change. For example, when the threshold is lowered to 5 consistent choices, slightly less than 43.5% of respondents who give random answers are classified. When the threshold is raised to 7, only about 2.5% of respondents who give random answers are still classified. Raising the threshold further makes only a minor difference. Applying an increased threshold of 7 consistent responses to our empirical sample would lead to an increase in unclassifiable responses from 11 out of 171 (6.4%, Table 2.3) to 29 out of 171 (17.0%). This gives the impression that many valid responses would not be classified at thresholds higher than 6.

Table 2.2. Percentage of random responses classified in each orientation on the 9-item TDM

Decision rule	*Prosocial*	*Individualistic*	*Competitive*	*Unclassified*
>= 5	14.485 %	14.485 %	14.485 %	56.55 %
>= 6	4.242 %	4.242 %	4.242 %	87.27 %
>= 7	0.828 %	0.828 %	0.828 %	97.52 %
>= 8	0.097 %	0.097 %	0.097 %	99.71 %
== 9	0.005 %	0.005 %	0.005 %	99.98 %

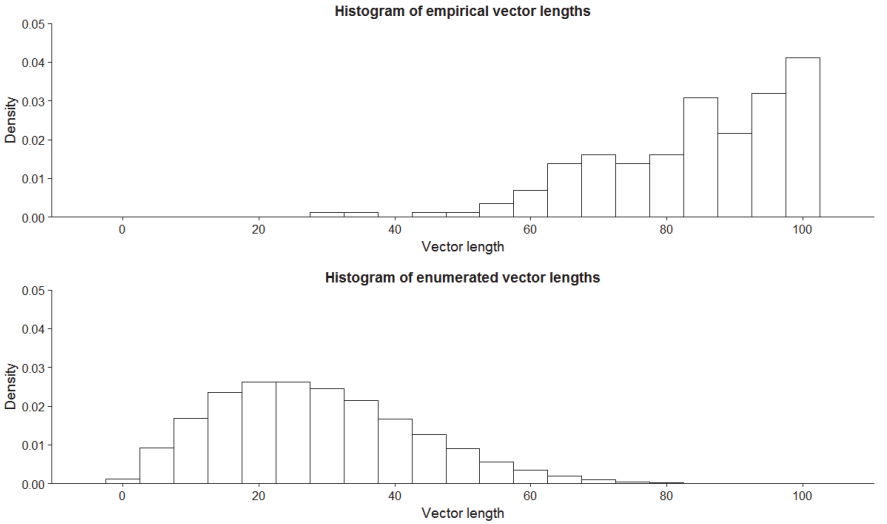
Ring measure

In order to determine the distribution of classifications which would be obtained for the ring measure if all respondents answered at random, we enumerated all possible decision profiles (i.e. all possible combinations of decisions) across the 24 items of the RM. Because the ring measure has two allocations to choose between (A and B) on each item, there are a total of $2^{24} = 16777216$ possible decision profiles. We

enumerated each decision profile (using the R programming language), then calculated the SVO angle and vector length for each decision profile. We classified each decision profile according to the angle boundaries specified in Table 2.1, with a minimum vector length of 25% of the maximum required in order to be classified (Au & Kwong, 2004).

Of the enumerated random decision profiles, 6.94% were classified as Altruistic, 13.81% as Cooperative, 13.88% as Individualistic, 13.81% as Competitive and 6.94% as Sadistic. The remaining 44.62% of random decision profiles did not satisfy the minimum vector length criterion and therefore remained unclassified. Based on the most commonly used threshold for classification (Au & Kwong, 2004), no less than 55.38% of participants who give entirely random answers are nonetheless classified into one of the SVO types. This suggests that the classification threshold of 25% of the maximum vector length may be too forgiving to effectively identify respondents who should not be classified. To investigate this further we can compare the distribution of vector lengths obtained from our enumeration of all possible random decision profiles to the distribution of vector lengths among the first observations of each respondent in our student sample. Figure 2.4 shows that there is very little overlap between the two distributions. In fact, if we were to move the classification threshold to 55.49% of the maximum vector length (the 95th percentile of the distribution of vector lengths from random answers) we would still classify 97.78% of our empirical cases. While it is possible that our sample is unusually consistent in their responses, so that a different empirical sample might contain more relatively low vector lengths, this result does suggest that there is room to exclude more false positives with a very minor increase in possible false negatives.

Figure 2.4. Empirical and enumerated distributions of vector lengths



Slider measure

The discrete slider measure consists of 6 primary items (we will ignore the secondary items for now), each of which has nine allocations to choose from. This means that there are 9^6 = 531441 possible decision profiles. We enumerated each of these decision profiles using R, then calculated the SVO angle for each decision profile. We classified each decision profile according to the angle boundaries specified in Table 2.1.

Of the enumerated random decision profiles, 0.05% were classified as Altruistic, 50.68% were classified as Cooperative, 49.23% were classified as Individualistic and 0.05% were classified as Competitive. This distribution of classifications among random answers is the most unbalanced among the three measures of SVO. The two categories which are underrepresented are also the least common in empirical samples. This means that a dataset which consists entirely of random responses might not be immediately distinguishable from a genuine sample.

We next applied the transitivity check to all enumerated random decision profiles. Murphy et al. (2011) state that SVO preferences should be transitive and that responding randomly would likely result in an intransitive set of responses. If so, this transitivity check could effectively function as a threshold for consistency which is used to exclude random responses. As it turns out, 41.85% of all enumerated random

decision profiles resulted in transitive preferences. Moreover, the distribution of classifications does not change dramatically when profiles with intransitive preferences are excluded (0.1% Altruistic, 52.94% Cooperative, 46.84% Individualistic, 0.1% Competitive). Thus, even after the transitivity check respondents making random errors are still directed away from the altruistic and competitive categories.

The effectiveness of the transitivity check as a consistency criterion is limited. On an aggregated level, the enumerated random responses show substantially fewer transitive preferences than the empirical samples we know of. Nearly all responses from our empirical sample pass the transitivity check (98.87%), and this is similar to the 95% reported by Murphy et al. (2011)). This suggests that when researchers find a much lower percentage of transitive preferences there is a reason to be suspicious about the quality of responses. However, based on these results, the transitivity check should not be used to filter out random responses or determine whether an individual response is genuine or random. For that, the false positive rate of 41.85% is too high.

CONVERGENT VALIDITY

Next, we investigate the convergent validity of the three measures, which is to say we investigate how often the same individual is classified into the same Social Value Orientation on multiple SVO measures. We first present an overview of the available literature on the convergent validity of the three measures, followed by an empirical investigation based on our own sample.

Literature review

The three main measures of SVO which we investigate in this study use roughly the same SVO types, with the exception that the RM and SLM divide prosocials into cooperators and altruists. The RM can in principle distinguish between multiple other orientations as well. The version used in this study also distinguishes the sadistic orientation (negative weight on others' payoffs, regardless of own payoffs). Au & Kwong (2004) present meta-analyses of classifications for the 9-item TDM and the RM. They find that in a meta-analysis of all studies involving the 9-item TDM (N=41) or the RM (*N*=15) the median percentage of individuals identified as cooperators is very similar between the 9-item TDM (46%) and the RM (45%). The same is true for the median percentage of competitive types identified (13.4% for 9-item TDM, 10% for the RM), but not for individualists. Studies using the 9-item TDM find a smaller median percentage of individualists (25%) than studies using the RM (35%). This difference

may in part be explained by the fact that the 9-item TDM results in a higher percentage of unclassified individuals (12%) than the RM (6%).

When Murphy et al. (2011) presented the SLM they also performed a comparison of this new measure to the 9-item TDM and the RM. They found that the three measures produced very similar results except that, as discussed, the 9-item TDM results in a higher number of unclassified individuals (10%, vs 1% for RM and 0% for SLM). Again, these unclassified individuals seem to be mainly ones who would be identified as individualists in the other measures (26.5% individualists in 9-item TDM, vs 40.5% in RM and 36.5% in SLM). A complete overview of the percentages allocated to each type in the studies discussed can be found in Appendix 1. Murphy et al. (2011) found that the 9-item TDM and the RM classified respondents the same way in 67% of cases, the SLM and TDM matched on 74% of cases, and the SLM and RM matched on 75% of cases.

Results

To assess convergent validity in our sample, we present the results obtained for the first observation from each participant. Across all four waves, 182 students participated at least once. We use the first observation from each of these 182 students to assess the properties of the three SVO measures. Students could decline to participate in parts of the survey, and students sometimes left some items blank without explicitly declining participation, so the number of observations obtained for each measure is not exactly 182. All in all, we have 171 complete measurements for the 9-item TDM, 175 for the RM, and 177 for the SLM.

Table 2.3. Classifications by category for each measure

	9-item TDM		Slider		Ring	
	N	*%*	*N*	*%*	*N*	*%*
Altruistic	-	-	0	0 %	1	0.6 %
Cooperative [a]	112	65.5 %	119	67.2 %	78	44.6 %
Individualistic	45	26.3 %	54	30.5 %	93	53.1 %
Competitive	3	1.8 %	2	1.1 %	2	1.1 %
Sadistic	-	-	-	-	1	0.6 %
Unclassified/mixed	11	6.4 %	2	1.1 %	0	0.0 %
Total	171	100 %	177	100 %	175	100 %

Note. [a] *For the 9-item TDM this represents the Prosocial category. This measure does not distinguish between altruism and cooperativeness*

The percentage of people classified in each category per measure is presented in Table 2.3. The classifications according to 9-item TDM and the slider measure are quite similar, while the ring measure shows a much greater percentage of individualists than the other two measures. All measures classify very few respondents as competitive types. The SLM and RM differentiate between altruistic and cooperative respondents. The RM identifies one respondent as altruistic, the SLM does not identify any respondent as an altruist. Because of the 9-item triple dominance measure's method of classification, in which respondents are only classified if they answer consistently on at least 6 out of the 9 items, some participants remain unclassified. This was the case for 11 respondents (6.4%).

Match between the three SVO measures

Next, we look at the extent to which the classifications provided by the three measures match. Presented below are cross-tables for each combination of measures (Tables 2.4, 2.5, and 2.6). Before we present these results it should be noted that the categories available for each measure are not entirely consistent. The SLM and RM both include an altruistic type, which is not included in the 9-item TDM. The RM includes a sadistic type, which is not included in the SLM or 9-item TDM. The 9-item TDM is the only one in which some respondents remain unclassified. As only one respondent was classified as altruistic (only on the RM), only one respondent was classified as sadistic (on the RM), and the number of unclassified respondents on the 9-item TDM is low (11 respondents, representing 6.4% of the sample), this does not present significant problems for the comparisons we are about to make.

Overall we find that the 9-item TDM and the SLM classify 74.3% of respondents the same way. Similarly, the SLM and the RM assign the same type to 72.5% of respondents. The mismatch between the 9-item TDM and the RM, however, is greater. These two measures only match on 63.5% of respondents. This mismatch occurs particularly among participants classified as prosocial on the 9-item TDM. No less than 40.4% of these participants are classified as individualists on the RM.

Table 2.4. Match between 9-item triple dominance measure and slider measure [a]

		Slider							
		Altruistic		*Cooperative*		*Individualistic*		*Competitive*	
9-TDM	*Cooperative* [b]	0	0.0%	92	55.1 %	18	10.8 %	0	0 %
	Individualistic	0	0.0%	13	7.8 %	30	18.0 %	0	0 %
	Competitive	0	0.0%	0	0 %	1	0.6 %	2	1.2 %
	Unclassified	0	0.0%	7	4.2 %	4	2.4 %	0	0 %

Note. [a] *Percentages are of the total, N = 167;* [b] *For the 9-item TDM this represents the Prosocial category. This measure does not distinguish between altruism and cooperativeness*

Table 2.5. Match between ring measure and 9-item triple dominance measure [a]

		9-TDM							
		Cooperative [b]		*Individualistic*		*Competitive*		*Unclassified*	
Ring	*Altruistic*	0	0.0 %	1	0.6 %	0	0.0 %	0	0.0 %
	Cooperative	65	38.9 %	5	3.0 %	0	0.0 %	4	2.4 %
	Individualistic	43	25.7 %	39	23.4 %	0	0.0 %	7	4.2 %
	Competitive	1	0.6 %	0	0.0 %	1	0.6 %	0	0.0 %
	Sadistic	0	0.0 %	0	0.0 %	1	0.6 %	0	0.0 %
	Unclassified	0	0.0 %	0	0.0 %	0	0.0 %	0	0.0 %

Note. [a] *Percentages are of the total, N = 167;* [b] *For the 9-item TDM this represents the Prosocial category. This measure does not distinguish between altruism and cooperativeness*

Table 2.6. Match between ring measure and slider measure [a]

		Slider							
		Altruistic		*Cooperative*		*Individualistic*		*Competitive*	
Ring	*Altruistic*	0	0.0 %	1	0.6 %	0	0.0 %	0	0.0 %
	Cooperative	0	0.0 %	74	43.3 %	3	1.8 %	0	0.0 %
	Individualistic	0	0.0 %	42	24.6 %	48	28.1 %	0	0.0 %
	Competitive	0	0.0 %	1	0.6 %	0	0.0 %	1	0.6 %
	Sadistic	0	0.0 %	0	0.0 %	0	0.0 %	1	0.6 %
	Unclassified	0	0.0 %	0	0.0 %	0	0.0 %	0	0.0 %

Note. [a] *Percentages are of the total, N = 171*

TEST-RETEST RELIABILITY

Next, we investigate the test-retest reliability of the three measures, which is to say we investigate how often the same individual is classified into the same Social Value Orientation across multiple repeated measurements. We first present an overview of the available literature on the test-retest reliability of the three measures, followed by an empirical investigation based on our own sample.

Literature review

Social value orientation is regarded as a trait (i.e. a property which is relatively stable over time) which reflects how people evaluate outcomes for self and others (Bogaert et al., 2008; Messick & McClintock, 1968). According to Bogaert et al. (2008), an often cited definition of SVO states that it 'reflects stable preferences for certain patterns of outcomes for oneself and others" (e.g. Van Lange et al., 1997). The evidence is mixed on whether responses to measures of Social Value Orientation are sensitive to contextual influences (Au & Kwong, 2004; Bogaert et al., 2008), and the number of studies is limited in this regard.

The 9-item TDM has been evaluated for stability at periods of up to nineteen months (see Bogaert et al. (2008) for an overview). Test-retest coefficients reported in Bogeart et al. (2008) for periods from 1 month to 19 months between tests range from 60% same classification to 75% same classification. Bogeart et al (2008) do not give more information on which individuals were likely to change in classification, or whether studies with longer times between tests were likely to have lower test-retest coefficients. Murphy et al. (2011) also report test-retest coefficients, based on a one-week time interval between tests. They found that 70% of individuals were classified the same way in both tests.

The RM has been evaluated for stability at periods of up to two months (Dehue, McClintock & Liebrand 1993 in Au & Kwong 2004). The reported test statistic was a Gamma (index of ordinal association) of 0.82. Murphy et al. (2011) report test-retest coefficients with two weeks between tests. They found that 68% of individuals were classified the same way in both tests. Murhpy et al. (2011) also report the correlation between SVO angles in the first measurement and the second measurement, which was 0.599.

The SLM is the most recently developed measure of the three and has only been tested for reliability by its designers (Murphy et al., 2011) with one week between tests. In this study, 89% of participants were classified the same way in both measurements. Additionally, the authors report the correlation between SVO angles in

the first measurement and the second measurement to be 0.915. The stability of these results is very high compared to the statistics reported for the 9-item TDM and the RM.

Results

We compare how classifications on each of the three measures changed between the first and second waves. In total, 89 respondents took part in both the first and second wave of data collection. Of these, 84 respondents completed the 9-item triple dominance measure at both time points. The slider measure and ring measure, respectively, have 86 and 83 complete measurements at both time points.

The percentage of respondents classified as the same type in the first and second wave is highest for the slider measure. Using this measure, 77.9% of respondents were classified the same way at both time points. The percentage of consistently classified respondents was lowest for the 9-item TDM, with exactly one-third of participants being classified in different categories in the first and second wave of the survey. The ring measure scores in between the other two measures, with 71.1% of respondents classified consistently as one type.

Across the three measures, between twenty percent and one-third of respondents were assigned a different SVO type in the first wave than in the second wave. [2] This raises the question which respondents changed classification from one wave to the next. A reliable and valid measure of Social Value Orientation should be able to detect large changes in a respondent's preferred allocation of payoffs between themselves and another person, but should not result in drastic changes as a result of small (random) fluctuations in a respondent's choices.

One way to investigate this for the slider measure and ring measure is to look at the size of changes in the raw angle rather than in the final classification. On the slider measure, the boundaries between types are approximately 35° apart, and a shift from an altruistic classification to a competitive classification would require a change in angle of more than 69°. On the ring measure, the boundaries between types are 45° apart, and a shift from an altruistic classification to a competitive classification would require a change in angle of more than 90°.

[2] We further compared classifications from Waves 3 and 4 to classifications from Wave 1. The more time has passed since the initial measurement, the greater the percentage of individuals whose classification has changed. This suggests some genuine change in SVO over time. However, these results must be interpreted with caution as only 27 individuals participated in both Wave 1 and Wave 4 while only 22 participated in all four waves. More detail on these descriptive analyses is available in an online appendix at https://osf.io/6rdx9/?view_only=87831a672837458eb667abe89bc818e1.

Table 2.7. Percentage of consistent classifications by measurement and SVO type [a]

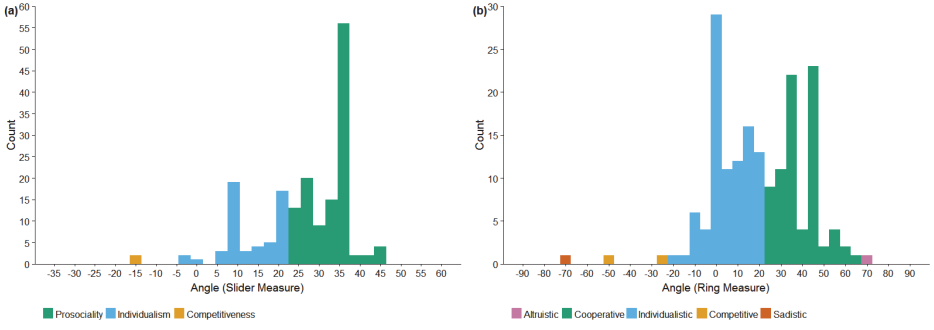
	9-item TDM		Slider		Ring	
	N (W1)	*% (W1)*	*N (W1)*	*% (W1)*	*N (W1)*	*% (W1)*
Cooperative [b]	58	74.1 %	55	85.5 %	37	67.6 %
Individualistic	22	54.5 %	30	66.7 %	44	75.0 %
Competitive	1	100.0 %	1	0.0 %	1	0.0 %
Sadistic	-	-	-	-	1	0.0 %
Unclassified/mixed	3	0.0 %	-	-	0	-
Total	84	66.7 %	86	77.9 %	83	69.9 %

Note. [a] *Based on SVO type in the 1st wave of data collection, compared to the 2nd wave 3 months later;* [b] *For the 9-item TDM this represents the Prosocial category. This measure does not distinguish between altruism and cooperativeness*

For the slider measure, the mean absolute change in angle between Wave 1 and Wave 2 was 7.56° (*SD* = 7.68°). The computed angle for approximately 85% of respondents changed less than 15° between Wave 1 and Wave 2. For the ring measure the mean absolute change in angle was somewhat higher (13.75°, *SD* = 18.78°). This was strongly influenced by a single extreme case, who changed from a sadistic orientation in Wave 1 to an altruistic orientation in Wave 2. If this case is left out, the mean absolute change in angle for the ring measure is 11.99° (*SD* = 9.65°). For this measure, approximately 75% of respondents saw a change in angle of less than 20°. Compared to the distances between the boundaries of a type and compared to the overall scale of the measures, the observed changes in angle are not large.

Figure 2.5: Histograms of (a) slider and (b) ring measure angles [a]

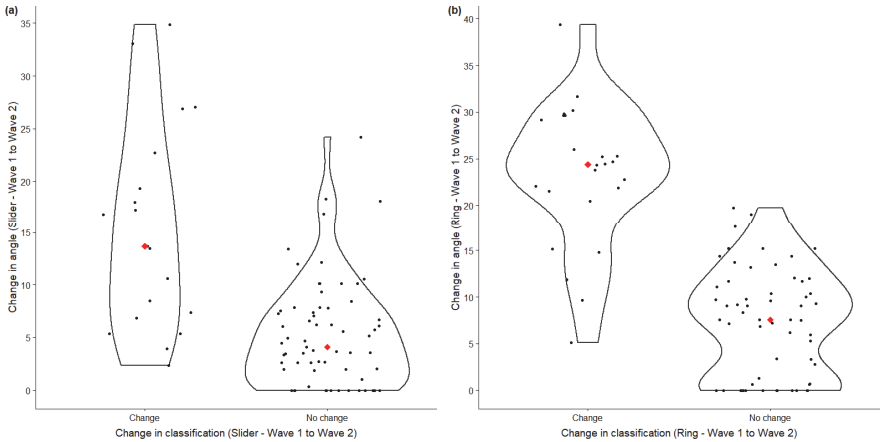
Note. [a] *Based on the first observation from each participant*



In fact, what we observe is that in Wave 1 many respondents are clustered relatively close to the boundaries of the type they were assigned. While there are clear groups of pure cooperators or pure individualists, there are also many respondents who perhaps lean towards one type but really have more nuanced preferences. This is illustrated, for example, by a histogram of respondents' computed SVO angles on the slider measure and the ring measure (Figure 2.5). While the distribution shows high peaks at the angles which correspond exactly to the midpoints of the SVO categories, the majority of respondents fall somewhere between pure cooperativeness and pure individualism. Many respondents are closer to the boundary with another type than they are to the typical angle of the type as which they were classified.

The majority of changes in classification on the slider measure and the ring measure are due to comparatively small changes in SVO angle. Figure 2.6 illustrates this observation. Although those respondents with the largest changes in angle did also change classifications, many respondents who changed classification show changes in the angle which also occur among those who remained classified the same way. This is particularly true for the slider measure (Figure 2.6 (a)). It appears that while both measures are able to detect large changes in angle, both measures also show changes in classification for a number of participants whose SVO angle changed only slightly. Many changes in classification are found among respondents whose SVO angle did not change all that much, but who were already positioned close to the boundary with another category. Reducing these measures to categorical classifications thus not only results in a loss of information on individual differences within those categories but also causes a loss of information on individual changes over time.

Figure 2.6: Absolute changes in angle on (a) the slider and (b) the ring measure, split by change in classification from Wave 1 to Wave 2 [a]



Note. [a] *Scores are jittered horizontally,* ◆ *indicates the median score. One extreme case (change in angle of 156.06°) is removed for the ring measure*

Analysis of continuous measures

The Ring Measure and the Slider Measure can be used as continuous measures (based on the computed SVO angles) rather than as categorical classifications. The Pearson correlation between the angle on the Ring Measure and the angle on the Slider Measure is 0.6916 among the 171 respondents who completed both the RM and the SLM. This is similar to the mean correlation between these measures reported by Murphy et al. (2011), which was 0.649.

Pearson correlations between angles in the first and second waves are 0.404 for the RM (83 complete observations) and 0.603 for the SLM (86 complete observations). The correlation between repeated measurements is higher for the SLM than for the RM, as reported by Murphy et al. (2011). Both correlations are lower than in Murphy et al.'s (2011) study, which is not unexpected given the longer interval (approximately 3 months versus 1 or 2 weeks). The difference between the two measures is also smaller. Murphy et al. (2011) reported a correlation of 0.915 for the SLM angles and 0.599 for the RM angles at intervals of 1 and 2 weeks respectively.

DISCUSSION

Conclusions

We have evaluated the three most commonly used measures of Social Value Orientation on several properties. First, we investigated the content validity of the three measures by examining how many respondents remain unclassified and how effectively the measures exclude random responses. Second, we investigated the convergent validity of the three measures. Third, we investigated the test-retest reliability of the three measures over a period of approximately three months. Fourth, we illustrated the differences between categorical and continuous measures of SVO. After this evaluation, the following conclusions can be drawn.

Regarding content validity, the three measures differ in their sensitivity to random responses. The 9-item TDM performs best in this regard, excluding the vast majority of random answers using the standard classification rule. The ring measure allows many random answers to be classified when using the most commonly used consistency threshold, but our results suggest that this threshold could be placed much higher to exclude more random answers while not excluding many cases from our empirical sample. The slider measure performs better than the ring measure when the suggested check for transitivity is applied, but still many random responses pass this check. Unlike the ring measure's consistency threshold, the slider measure's transitivity is binary and thus cannot be adjusted to exclude more random responses. In a sample with many random responses, the slider measure can result in many unjustified classifications. What is more, the mistake may well go undetected when researchers only look at the distribution of classifications, since these random responses tend to be classified in the categories which also contain most genuine responses. The transitivity check likely will identify samples with unusually many random answers.

Having said that, we should note that in most studies it is unlikely that a significant proportion of respondents will respond entirely at random. We know, for example, that the 9-item TDM generally results in around 12% of respondents remaining unclassified (Au & Kwong, 2004) while a much larger percentage would be expected if a significant proportion of respondents were answering at random.

It should also be noted that it may be unrealistic to expect respondents who do not take their response seriously to answer at random. Rather, we might expect that these respondents choose some low effort method of completing the questionnaire such as always choosing the first option, or the last option, or some other predictable pattern. We have not included a comprehensive evaluation of the performance of

each measure against such response patterns because the outcomes may depend on details of the questionnaire being used (such as the order of questions and items). An overview of the predictable patterns we applied to each measure is available in an online appendix (see footnote 1).

Regarding convergent validity, the classifications obtained by the ring measure seem to differ substantially from those obtained by the other two measures. The ring measure identifies more individualists than the other two measures. This finding is consistent with previous research (Au & Kwong, 2004; Murphy et al., 2011) and is worth exploring further. We also find that particularly between the 9-item TDM and the ring measure, many respondents are not classified the same way. Nonetheless, the three measures are about as consistent with each other as they are with themselves after a three month period, which again matches previous research (Murphy et al., 2011).

Regarding test-retest reliability, the slider measure shows the least change in classifications after a three month period (77.9% consistency or a correlation between angles of $r = 0.603$). Given the substantial evidence that Social Value Orientation is a stable personality trait (see Bogaert, Boone, & Declerck, 2008, for an overview) this is a desirable property. However, the test-retest reliability of the Slider Measure is not as high as in the only other evaluation we know of (89% consistency or a correlation between angles of $r = 0.915$), in which test and retest were one week apart (Murphy et al., 2011). We have not reported extensively on changes in classification over periods longer than three months due to the small sample size in these comparisons (n = 27 for comparison of Wave 1 to Wave 4). However, it may be worth noting that the consistency of classifications after roughly a year and a half (Wave 1 to Wave 4, see footnote 1) mirrors the pattern observed after three months. The SLM is the most consistent (66.7 %), followed by the RM (60%), followed by the 9-item TDM (52%). The results suggest that the SLM consistently has higher test-retest reliability than the 9-item TDM and the RM.

Finally, we want to devote some attention to the choice between a categorical or a continuous representation of Social Value Orientation. Recent literature on SVO frequently mentions that SVO is a continuous concept which should not be reduced to categories (e.g. Bogaert et al., 2008; Murphy & Ackermann, 2012; Murphy et al., 2011; Pletzer et al., 2018). Conceptually, given that SVO is defined in terms of the balance between the weight an individual attaches to their own outcomes and the weight they attach to the outcomes of another, a continuous representation of SVO makes sense. Additionally, Murphy et al. (2011) show that there is significant variation in SVO angles

within the traditional categorical orientations. Adding further evidence for the benefits of a continuous measure of Social Value Orientation, we find that much of the change in classifications observed after a three month period may be due to relatively small changes in respondents' answers. These small changes may be just sufficient to shift the respondent into a different classification. This gives the appearance of a much more drastic shift in this respondent's orientation than is actually expressed in the respondent's answers.

Limitations

Our study has several limitations related to our sample and our opportunities for data collection. First, our sample is very homogeneous, consisting entirely of sociology students. This may have affected the distribution of orientations we observed as well as the consistency of answers. Although the distribution of orientations we have obtained is not inconsistent with previous research, including research using psychology students (Van Lange et al., 2011) and research using a representative sample of the Dutch population (Van Lange et al., 1997), samples from different fields (Van Lange et al., 2011) or different countries may show very different distributions. Additionally, the sample suffered from attrition, particularly (unfortunately) among the group of students which we tracked through all four waves of the study. This results in a small sample size (only 27 students) when comparing the first to the last wave. As a result, we have relegated analyses of test-retest reliability over periods longer than three months to an online appendix (see footnote 1).

The SVO questionnaires in our study were not incentivized. That is to say, the payoff allocations our respondents chose were purely hypothetical and no money was distributed to our respondents. Offering incentives, or making the study incentive compatible by using something like a lottery, may be preferable. Incentivizing such measures is believed to improve the honesty of responses (Murphy & Ackermann, 2012). The non-incentivized nature of our questionnaire may have biased our results either by causing our respondents to respond more prosocially than they otherwise would (since monetary incentives favor individualistic orientations) or by making our respondents take the questionnaires less seriously. The fact that our respondents' choices appear to be very consistent mitigates the concern that they may not have taken the questionnaires seriously. Whether, or to what extent, reported SVO is affected by incentives is unclear. As far as we are aware there are no studies directly comparing incentivized and non-incentivized SVO questionnaires. Evidence on the *effects* of SVO in incentivized and non-incentivized context is mixed. For example,

Balliet et al. (2009) report that the effect of SVO on cooperation is larger in non-incentivized social dilemmas than in incentivized social dilemmas, while Pletzer et al. (2018) find that the effect of SVO on the expectation of others' cooperation is the same in incentivized and in non-incentivized social dilemmas. Overall, we do not have significant doubts about the accuracy of our results in particular and the results of non-incentivized measures of SVO in general.

Recommendations

Based on our results and on previous literature on the advantages and disadvantages of various measures of Social Value Orientation (Murphy & Ackermann, 2012) we come to the following recommendations. First, we concur with recent remarks in the literature that it seems worthwhile to attempt to avoid SVO classifications and instead use a continuous measure of SVO (such as the angle computed by the ring measure or the slider measure). Restricting Social Value Orientations to a categorical measure is not only inconsistent with the definition of SVO, but it also discards valuable information on inter- and intrapersonal variation (see also Murphy et al., 2011) and overestimates changes in SVO over time.

Second, based on our results and on the other relevant properties of the various measures as discussed in the literature, we recommend the Slider Measure as the most suitable measure of SVO for most situations. From the literature we know that Slider Measure is efficient and easy to implement (particularly compared to the RM), allows for a continuous measure of SVO, allows a distinction between inequality aversion and joint outcome maximization, and produces internally consistent results (Murphy & Ackermann, 2012; Murphy et al., 2011). We can confirm the observation made by Murphy et al. (2011) that the Slider Measure has the highest test-retest reliability among the three measures tested and scores well on convergence with the other two measures. The Slider Measure mainly distinguishes itself from the 9-item TDM by enabling continuous measurement of SVO, and from the RM by being more efficient and more consistent.

We would, however, like to qualify this recommendation somewhat. For one, there are a number of properties of the Slider Measure which we have not ourselves evaluated (such as its internal consistency and its ability to predict behavior in social dilemmas) and for which we are relying on an evaluation by the original author of this measure. A more extensive independent review of the Slider Measure would be useful if this measure is to become the new standard in the field. For another, while we recommend the use of the Slider Measure in general, there may be situations in which

this measure should be used with caution. Particularly when a significant proportion of respondents can be expected to answer at random or at least not according to their true preferences (e.g. you are working with a sample which does not take the study seriously), researchers should be aware that the transitivity of respondents' preferences can be an indicator of whether many responses in a sample are not genuine (Murphy et al., 2011) but is not suited to determining the validity of an individual response.