Article:
Ranyard, R. orcid.org/0000-0003-4293-9510, Montgomery, H., Luckman, A. et al. (1 more author) (2024) Violations of transitive preference: A comparison of compensatory and noncompensatory accounts. Psychological Review, 131 (6). pp. 1392-1410. ISSN: 0033-295X

https://doi.org/10.1037/rev0000502

Violations of transitive preference:
A comparison of compensatory and noncompensatory accounts

Rob Ranyard, [1], Henry Montgomery, [2] Ashley Luckman[3] and Emmanouil Konstantinidis[4]

[1]Centre for Decision Research, University of Leeds, UK

[2]Department of Psychology, Uppsala University, Sweden

[3]University of Exeter Business School, University of Exeter, UK

[4] Department of Psychology, University of Warwick, UK

Author Note

Correspondence concerning this article should be addressed to Rob Ranyard, Centre for Decision Research, Leeds University Business School, Moorland Rd, Leeds, West Yorkshire, LS6 1AN, UK. Email: r.ranyard@leeds.ac.uk. A preliminary outline was presented online at the SPUDM conference, 22-24 August 2021, University of Warwick, UK.

We are very grateful to Daniel Cavagnaro and Clintin Davis-Stober, and to Tobias Kalenscher, for providing us with the data from their studies, and to Peter Ayton for valuable feedback on an earlier version. All data and analysis code for the new analyses reported here can be accessed at: https://osf.io/9c4nu/

Abstract

Violations of transitive preference can be accounted for by both the noncompensatory lexicographic semiorder heuristic and the compensatory additive difference model. However, the two have not been directly compared. Here we fully develop a simplified additive difference (SAD) model, which includes a graphical analysis of precisely which parameter values are consistent with adherence to, or violation of, transitive preference, as specified by weak stochastic transitivity (WST) and triangle inequalities (TI). The model is compatible with compensatory, within-dimension evaluation. We also develop a stochastic difference threshold (SDT) model which also predicts intransitive preferences and encompasses a stochastic lexicographic semiorder model. We apply frequentist methods to compare the goodness of fit of both of these models to Tversky's (1969) data and four replications, and Bayes factor methods to determine the strength of evidence for each model. We find that the two methods of analysis converge and that, for two-thirds of the participants for whom predictions can be made, one of these models predicting violations of WST has a good, and the best, fit, and has strong Bayesian support relative to an encompassing model. Furthermore, for about twenty percent of all participants the SAD model (consistent with violations of WST or TI) is significantly better-fitting and has stronger Bayesian support than the SDT model. Finally, Bayes factor analysis finds strong evidence against transitive models for most participants for whom the SAD model consistent with violation of WST or TI is strongly supported.

Keywords: Risky choice; weak stochastic transitivity; triangle inequalities; simplified additive difference model; stochastic difference threshold model; lexicographic semiorder model; violation of transitive preference

Violations of transitive preference: A comparison of compensatory and noncompensatory accounts

The transitivity of preferences, the axiom that if A is preferred to B and B to C then A is preferred to C, is a cornerstone of rational models of decision making (Savage, 1954; von Neumann & Morgenstern, 1944). However, the investigation of whether this is a property of human preferential choice is complicated by the fact that when people are presented with the same choice on multiple occasions, they do not always choose the same alternative. That is, preferential choice is probabilistic (Mosteller & Nogee, 1951; Rieskamp, 2008). To accommodate this, the basic research question has become: does preferential choice adhere to, or violate, probabilistic specifications of transitivity? Here we focus on two differing ways in which probabilistic specifications of transitivity have been operationalized: weak stochastic transitivity (WST) as investigated by Tversky (1969); and the mixture model of transitive preference[1], analyzed by Regenwetter, Dana & Davis-Stober (2010, 2011). WST is satisfied when, for all a, b, c from a set of alternatives,

$$p(a, b) \geq .5 \text{ and } p(b, c) \geq .5 \text{ implies } p(a, c) \geq .5,$$

where p(x, y) represents the probability of choosing x on presentation of alternatives (x, y). The mixture model of transitive preference is, for up to five alternatives, equivalent to the so-called triangle inequalities (TI) condition (cf. Marschak, 1960),

$$p(a, b) + p(b, c) - p(a, c) \leq 1,$$

for all a, b, c from a set of alternatives.

We contribute to this fundamental question for the understanding of human decision making in several ways. First, unlike previous studies, we evaluate both compensatory and noncompensatory dimension-based accounts of violations of transitivity. Some previous

[1] Both WST and the mixture model of transitive preference have their origins in the late 1950s and early 1960s, in the work of Block and Marschak (1955) and others; see also Loomes and Sugden (1995), who describe the mixture model of transitive preference as the random preference model, and Regenwetter et al. (2010, 2011) for further details.

studies have focused wholly on transitive models, which, by definition, cannot account for violations (Cavagnaro & Davis-Stober, 2014; Regenwetter, Dana & Davis-Stober, 2010, 2011). Others have investigated the extent to which either the noncompensatory lexicographic semiorder model (Birnbaum & Gutierrez, 2007; Davis-Stober, Brown & Cavagnaro, 2015) or the compensatory additive difference model (Kalenscher et al., 2010; Ranyard, Montgomery, Konstantinidis & Taylor, 2020) can account for violations of transitivity. However, few have compared the two (but see Montgomery, 1977). Here, in order to compare these alternative accounts, we: (1) devise a graphical analysis for the simplified additive difference (SAD) model (Ranyard et al., 2020) which identifies precisely the parameter values consistent with adherence to, or violation of, transitive preference as specified by both WST and TI; and (2) compare the SAD model to a new stochastic difference threshold (SDT) model, which encompasses a stochastic lexicographic semiorder model. We fully investigate violations of both of the above probabilistic specifications of transitivity (WST and TI) by examining the extent to which violations of, as well as adherence to, these models are consistent with the SAD or SDT model. In addition, we extend Tversky's (1969) test of predictions of WST violation from a pretest to two additional data sets. Finally, we analyze violations of, and adherence to, WST and TI by using two different analytic methods: classical frequentist tests of model fit; and Bayes factor strength of evidence techniques. These analyses are supported by the graphical analysis which gives a visual representation of the correspondence between the SAD model's description of violations and adherences to WST and TI, and observed choice proportion's violations and adherences to WST and TI. In the above ways we offer the most complete analysis yet undertaken of alternative dimension-based accounts of violations of transitive preference.

Tversky's (1969) dimension-based models

From the 1950s, economists and psychologists conjectured whether there are circumstances in which preferences might be systematically intransitive (Edwards, 1954; May, 1954). However, Tversky (1969) was the first to present evidence of intransitive preferences that were both systematic and predictable, and could not be explained away as errors on the part of study participants. Developing the ideas of May, Tversky argued that intransitive preferences might occur when people construe decision alternatives as varying on two or more dimensions (such as lotteries which differ in outcome probability and outcome amount) and process information within dimensions rather than within alternatives. He proposed two different dimension-based models to account for intransitive preferences: the above-mentioned lexicographic semiorder heuristic and the additive difference model.

The lexicographic semiorder heuristic is a noncompensatory process model in which only some of the available information is processed in some contexts. For two-dimensional alternatives, Tversky (1969, p. 32) defined it as follows: "if the difference between the two alternatives on dimension I is (strictly) greater than [a threshold value,] ε, choose the alternative that has the higher value on dimension I. If the difference between the alternatives is less than or equal to ε, choose the alternative that has the higher value on dimension II". This noncompensatory model can explain violations of transitive preference in a set of alternatives where some dimension differences are below the threshold and others are above it. Tversky investigated this prediction in an experiment involving choice between simple monetary lotteries with the structure win s dollars with probability p, otherwise win zero (dimensions S and P). The five lotteries he used are shown in Table 1, labelled a, b, c d e. According to the lexicographic semiorder heuristic, if the win probability, P, is dimension I, people will prefer the lottery with the better P-value when the P-dimension difference between them is greater than a threshold (say, 2/24 in Tversky's lottery set). However, they will switch to a preference for the lottery with the better S-value (winning amount) if the P-dimension difference is less than or equal to the threshold. In this case some preferences will violate transitivity, for example, a ≻ b, b ≻ d, but d ≻ a (where x ≻ y denotes x is preferred to y). Tversky defined preference stochastically, such that if x ≻ y then p(x, y) > .5, and his specific prediction was that participants' choices would violate WST.

Tversky's (1969) second model to account for violations of transitivity, the additive difference model, fully utilizes all available information by evaluating and comparing all dimension differences. This model, fully described later, is a mathematical, as-if, model that is compatible with a compensatory process of weighing subjective dimension differences, the advantages and disadvantages of alternatives, against each other. In the case of simple lotteries, the difference between the alternatives in win probability, P, is weighed against the difference in winning amounts, S. Tversky proved that preferences are transitive under this model only in certain restricted cases, including where all alternatives are one-dimensional. He did not, however, specify any subjective difference functions that could account precisely for violations of transitivity.

Tversky's seminal work on this issue left many open questions to which he did not return in his subsequent research. The final words of his paper were: "The main interest in the present results lies not so much in the fact that transitivity can be violated but rather in what it reveals about the choice mechanism and the approximation method that govern preference between multidimensional alternatives" (1969 p. 46). He did not, in subsequent empirical studies, return to his 1969 lottery paradigm[2]. Neither did he further develop the implications of the additive difference model, even though both of these dimension-based models were consistent with his results. Rather, the focus of his discussion of the findings, both in the original paper and later, was in terms of the noncompensatory lexicographic semiorder

[2] Nor did he return to the second study of his 1969 paper that presented hypothetical choices between job applicants varying in three dimensions, intellectual ability, emotional stability and social facility.

heuristic as an approximation method that could explain intransitive preferences. Consistent with this account, Kahneman and Tversky (1979) interpreted the 1969 findings as evidence of prospect theory's editing phase; specifically, the editing operation whereby small differences are eliminated as part of the simplification of decision alternatives prior to their evaluation. This editing process was retained in their subsequent development of cumulative prospect theory (Tversky & Kahneman, 1992). In later work on preference reversals, with Slovic and others, Tversky argued that the lexicographic semiorder heuristic account was consistent with later evidence that decision makers tend to rely on evaluating the prominent dimension in multidimensional decisions (Lichtenstein & Slovic, 2006; Montgomery, Selart, Gärling, & Lindberg, 1994; Slovic, 1995; Tversky, Sattath & Slovic, 1988).

While the lexicographic semiorder account of intransitive preferences is consistent with prospect theory, any evidence supporting a compensatory dimension-based account would present a significant difficulty for this theory. The evaluation phases of both the original prospect theory and cumulative prospect theory predict transitive preferences. If the evidence reviewed here supports an additive difference model, then the intransitive preferences in Tversky's (1969) lottery context cannot be interpreted in terms of an editing operation with rather minor theoretical importance. Rather, it would be consistent with a compensatory process weighing dimension differences against each other, which is at variance with prospect theory's within-alternative evaluation function. It is important, then, to evaluate these two accounts of intransitive preference, which we do here.

Evidence of violations of transitive preference: Tversky's (1969) lottery study

As mentioned earlier, Tversky (1969) investigated the above-described prediction of the lexicographic semiorder heuristic in an experiment involving choice between pairs of lotteries from the set shown in Table 1. In the first session of the experiment ($N = 18$), the pretest, student participants chose between adjacent lotteries from the set, i.e., pairs (a, b), (b, c), (c, d) and (d, e) as well as the extreme pair, (a, e). Each of these was presented three times with filler pairs interspersed. Participants were told that, after the session, they would play one of the lotteries they had chosen on a randomly selected trial for real to determine their only session payment. Then those (eight) participants who had chosen the higher S on the majority of adjacent pairs at least twice, but the higher P on the extreme pair at least twice, returned for the main experiment. Tversky predicted that the preferences of these participants ($n = 8$) would violate WST in the second stage of the experiment. In this stage, participants chose from each of the ten pairs of the lottery set 20 times across five sessions, one week apart. As before, filler choices were interspersed and at the end of a session participants played one chosen lottery for real.

Tversky tested the goodness of fit of WST via frequentist likelihood ratio (LR) tests and found that the choices of five of these eight participants significantly violated WST at $p < .05$. He did not assess the transitivity of the other ten participants in his study as these were not predicted to be intransitive. His main finding, then, was that based on a pretest, 8/18 participants were predicted to violate WST, and when tested, 5/8 did so. The choice proportions of three of Tversky's participants predicted to violate WST are reproduced in the middle panels of Table 1. In each of these it can be seen that, descriptively, choice proportions violate WST for several triads of lotteries. This basic finding has been replicated several times over the years (Budescu & Weiss, 1987; Cavagnaro & Davis-Stober, 2014; Kalenscher et al., 2010; Montgomery, 1977; Ranyard, 1977). However, the evidence that WST is significantly violated by 5/8 of Tversky's participants predicted to do so has been disputed. Iverson and Falmagne (1985) identified an error in Tversky's frequentist LR test of WST, and when they applied a valid, inequality-constrained LR test, they found that WST was a poor fit to the data of only one of the eight participants[3]. On the other hand, Ranyard et al. (2020) found that the SAD model consistent with violation of WST was a good fit to the data for 6/8 of Tversky's participants. So, for five of the six participants, for each of whom Tversky observed choice proportions violating WST, both a model consistent with adherence to WST (Iverson and Falmagne, 1985), and one consistent with violation (Ranyard et al., 2020), are well-fitting models.

Regenwetter et al. (2011) argued that before concluding that preferences are intransitive one should investigate other probabilistic specifications of transitivity, in particular, the mixture model of transitive preference, which as mentioned earlier is equivalent to TI for five or fewer alternatives. These authors applied their frequentist inequality-constrained test of the goodness-of-fit of TI to Tversky's data and found that this probabilistic specification of transitivity was a good fit for at least six participants and a poor fit for at most two participants; the rest of the participants did not violate TI.

In a later analysis, Cavagnaro and Davis-Stober (2014) computed Bayes factors to assess the strength of evidence for and against both WST and TI for Tversky's data. Comparing each of these transitive models to an unconstrained model that allows violations of transitivity, they found strong evidence against WST for six of Tversky's participants, and strong evidence against TI for three of them (their choice proportions are shown in Table 1).

Overall, then, there exists evidence of violations of transitive preference in Tversky's original lottery study. However, it is somewhat incomplete. For one thing, Iverson and Falmagne's (1985) conservative goodness of fit analysis of WST has not been corroborated by the subsequently developed more accurate methods of Regenwetter, Davis-Stober and their colleagues. Although these methods have been applied to some replications of

[3] We show later that the subsequently developed more accurate methods find that WST was significantly violated by 3/8 participants.

Tversky's study (Regenwetter et al., 2010; 2011), their application to Tversky's original data, and several other published replications, is not on the published record. The evidence is incomplete in other respect too, as explained shortly.

Evidence of violations of transitive preference: replications of Tversky's lottery study

In this paper we focus on Tversky's (1969) lottery experiment and four of the six replications we analyzed in our earlier study[4]: Cavagnaro & Davis-Stober, (2014), Kalenscher et al. (2010), Montgomery (1977) and Regenwetter et al. (2011). We selected these data sets for review and reanalysis here because they all adopted the lottery set that Tversky devised (adjusted for currency and inflation) and had sufficient replications of each lottery pair to enable the planned analysis to be conducted.

The available evidence of violations of transitive preference in these replications can be summarized as follows. With respect to the frequentist analysis of WST, Regenwetter et al. (2010) found that the model was a good fit for all participants in their replication. However, Ranyard et al. (2020) found that the SAD model consistent with *violation* of WST was a good fit for four of these participants. With respect to TI, in addition to Tversky's study, Regenwetter et al. (2011) tested its goodness of fit for two of the replications reviewed here, namely Montgomery (1977) and the cash I data set of Regenwetter et al. (2011). Summarizing their results, we calculate that across the three studies the model was a poor fit for 9% of participants ($p < .02$), with low p-values ($p < .15$) for a further 16% of them. Turning to Bayesian analyses, Cavagnaro and Davis-Stober (2104) computed Bayes factors for WST and TI, in comparison to an encompassing model, for three of the above replications, while Brown et al. (2015) did so for the other (Kalenscher et al, 2010). Across

[4] In the present review we have excluded Tsai and Böckenholt's (2006) study because only the first four lotteries of Tversky's set were used and it was the only study that did not have a real consequence of a lottery choice. We also dropped Ranyard's (1977) study because it involved less than ten replications and a rather low sample size. One other study met our criteria, by Budescu and Weiss (1987), but unfortunately the choice proportion data for this was not available.

all these studies: (1) for over twenty percent of participants there is strong Bayes factor evidence against WST when compared to the unconstrained model (BF < .3); and (2) for about one quarter of participants there is strong evidence against TI.

Aims and objectives

Taken together, previous frequentist and Bayesian analyses show that violation of transitive preference is a phenomenon in need of an explanation. As explained at the beginning, in this study we comprehensively compare the validity of the noncompensatory lexicographic semiorder heuristic and the compensatory additive difference accounts in a reanalysis of Tversky's (1969) lottery experiment and four published replications. This requires new theoretical developments, presented in the next section. First, we present the SAD model together with a new graphical analysis of how it relates to violations of, and adherence to, WST and TI. Second, we present the new SDT model, which encompasses a stochastic lexicographic semiorder model.

In the two following sections we present our methods of analysis and our findings. The main aim of the frequentist analysis is to identify for each participant the good and best-fitting model, from a set of nested models, and whether this is consistent with adherence to, or violations of, probabilistic specifications of transitivity (WST and TI). Two secondary aims are to investigate the extent to which the SAD model successfully predicts violations of WST from pretests, and for Tversky's data only, to compare the goodness of fit of dimensional models consistent with violation of transitivity (SAD and SDT) with transitive models (WST and TI). The latter involves the analysis of WST using the methods of Regenwetter et al. (2014) and Zwilling et al. (2019) not reported in previous studies.

The aims of the Bayesian analyses are to compare the strength of Bayes factor evidence for the SAD and SDT models against each other, and to compare more broadly the Bayesian strength of evidence for dimensional models consistent with violations of transitivity (SAD and SDT) against that for transitive models (WST and TI). Finally, the main aim of the graphical analysis is to describe in detail the relationship between the best-fitting SAD model's consistency with violation of, or adherence to, WST and TI, and that of the observed choice proportions for each participant. This provides additional, detailed information on the performance of the SAD model, especially with respect to how it accounts for observed violations of transitive preference. In the final section we conclude with a discussion of the implications of these analyses for decision research, and document some open questions still to be addressed.

Table 1

Tversky's (1969) lottery, the observed choice proportions (cps) of three participants, and the maximum likelihood estimates (MLEs) of choice probabilities for the SAD model (r = 20 presentations of each lottery pair)

Lotteries			Tversky's (1969) participants: 1 (top), 3 (middle), 6 (bottom)									
			Observed cp					MLE				
P	S	Label	a	b	c	d	e	a	b	c	d	e
7/24	5.00	a	-	.75	.70	.45	.15	-	.823	.637	.397	.198
8/24	4.75	b		-	.85	.65	.40		-	.823	.637	.397
9/24	4.50	c			-	.80	.60			-	.823	.637
10/24	4.25	d				-	.85				-	.823
11/24	4.00	e					-					-
			Observed cp					MLE				
P	S	Label	a	b	c	d	e	a	b	c	d	e
7/24	5.00	a	-	.75	.70	.60	.25	-	.875	.721	.488	.261
8/24	4.75	b		-	.80	.65	.40		-	.875	.721	.488
9/24	4.50	c			-	.95	.80			-	.875	.721
10/24	4.25	d				-	1.00				-	.875
11/24	4.00	e					-					-
			Observed cp					MLE				
P	S	Label	a	b	c	d	e	a	b	c	d	e

7/24	5.00	a	-	1.00	.90	.65	.20	-	.883	.746	.533	.308
8/24	4.75	b		-	.80	.75	.55		-	.883	.746	.533
9/24	4.50	c			-	.90	.65			-	.883	.746
10/24	4.25	d				-	.75				-	.883
11/24	4.00	e					-					-

Notes: *P* is the win probability of the lottery, otherwise win nothing; *S* is the payoff in USD; MLE is the maximum likelihood estimate.

Theory

The simplified additive difference (SAD) model

We first specify the algebraic SAD model for Tversky's (1969) lottery paradigm and explain the conditions under which, depending on its parameter values, it describes transitive or intransitive preferences. We then present the stochastic SAD model[5] and its relationship to the WST and TI conditions. Finally, we consider models nested within the SAD model whose goodness of fit can be compared.

The algebraic SAD model

Tversky's (1969) lottery paradigm comprises a set of two-dimensional lottery alternatives, A = S x P, of the form (s_i, p_i) such that payoff s_i is won with probability, p_i, otherwise nothing is won, i = 1, … 5. The set of payoffs, s_1, …., s_5 are decreasing in equal intervals, d_s, and the set of probabilities, p_1, …. , p_5 are increasing in equal intervals, d_p. We also denote the five lotteries of set A as a = (s_1, p_1), b = (s_2, p_2), c = (s_3, p_3), d = (s_4, p_4), e = (s_5, p_5). The specific values of Tversky's original lottery set are shown in Table 1. For this lottery set the additive difference model assumes that there are scales $u_s(s_i)$ and $u_s(p_i)$ which represent the subjective values of payoffs on dimension S, and probabilities on dimension P, respectively (i = 1, …, 5). The difference between the subjective values on the S and P dimensions are denoted δ_s and δ_p respectively. The additive difference model also specifies the functions $\varphi_s(\delta_s)$ and $\varphi_p(\delta_p)$, which are the subjective values of the differences between the subjective values on each dimension. These can be viewed as the perceived advantages or disadvantages of lottery x over y on the P and S dimensions. According to Tversky, $\varphi_p(-\delta_p) = -\varphi_p(\delta_p)$ and $\varphi_s(-\delta_s) = -\varphi_s(\delta_s)$. In the context of A = S x P, Tversky's algebraic additive difference model states that:

(1) $x \succcurlyeq y$ if and only if $\varphi_s(\delta_s) - \varphi_p(\delta_p) \geq 0$

[5] Subsequently, we refer to the extended (stochastic) SAD model as the SAD model, and its algebraic component as the algebraic SAD model.

where x has a lower P and higher S value than y in Tversky's lottery set, δ_s is the difference between the subjective values of the payoffs of x and y, and δ_p is the difference between the subjective values of the win probabilities of x and y.

To specify the SAD model we simplify Tversky's additive difference model as follows. First, we assume that the subjective values of the payoffs and win probabilities of the lotteries are equal to their objective values, i.e., $u_s(s) = s_i$ and $u_p(p) = p_i$. Second, we assume that the differences in subjective values of the payoffs (dimension S) and of the win probabilities (P dimension) are the differences between their objective values, standardized by the objective difference between them, i.e., $\delta_p = (p_i - p_j)/d_p$ and $\delta_s = (s_i - s_j)/d_s$. This standardizes the differences in subjective dimension values, δ_p and δ_s, to a common scale, the objective standard difference level between lotteries of the set, d_c. That is, $\delta_p = \delta_s = d_c$, which for Tversky's lottery set takes the values 1, 2, 3 or 4. It follows from these simplifications that the subjective value functions, φ_p and φ_s, of the differences between the subjective values on each dimension, are also functions of d_c. The algebraic additive difference model becomes:

(2) $x \succcurlyeq y$ if and only if $\varphi_s(d_c) - \varphi_p(d_c) \geq 0$

where x has a lower P and higher S value than y in Tversky's lottery set, and d_c is the objective standard difference between x and y.

If either $\varphi_s(d_c)$ or $\varphi_p(d_c)$ is nonlinear and the other is linear, or if both are linear but not related by $\varphi_s(d_c) = \varphi_p(kd_c)$ for some positive k, then Tversky's (1969) transitivity condition is not met and preferences may be intransitive under the model.

The algebraic SAD model is completed by one further simplification. The *overall* subjective difference function, $osd(d_c) = \varphi_s(d_c) - \varphi_p(d_c)$ is modelled (rather than separately modelling each subjective dimension difference function, $\varphi_s(d_c)$ and $\varphi_p(d_c)$). If the overall subjective difference function, $osd(d_c)$, is either monotone increasing or monotone decreasing, and changes from positive to negative, or vice versa, in the range $1 < d_c < 4$, then preferences will be intransitive. Any overall difference function with the above properties, either curvilinear, linear, or a step function, would be consistent with intransitive preferences. However, a linear function is the most parsimonious, and at the same time is compatible with a compensatory process of weighing advantages and disadvantages against each other. For these reasons we complete the specification of the algebraic SAD model for Tversky's lottery paradigm by a two-parameter linear function:

(3) $osd(d_c) = a_0 + a_1 d_c$

where d_c is the objective dimension difference level between lotteries x and y described earlier.

Condition 2 and Equation 3 together provide the full specification of the algebraic SAD model.

The algebraic SAD model and intransitive preferences

Identifying when preferences will be transitive or intransitive in Tversky's lottery paradigm under the algebraic SAD model is straightforward. This is illustrated in Figure 1, which shows the graphs of four specifications of the model. In two of them the linear function crosses the horizontal axis, osd(0), at $d_c = -a_0/a_1$, between $d_c = 1$ and $d_c = 4$. In these cases, preference switches from the better S to the better P, or vice versa, in the range $1 < d_c < 4$, and some preference cycles will be intransitive; for example, for the functions with a negative slope that cross the horizontal axis, $osd(d_c) > 0$ for pairs (a, b) and (b, d), but $osd(d_c) < 0$ for pair (a, d) in the lotteries of Tversky's paradigm, resulting in the cycle $a \succ b$, $b \succ d$ and $d \succ a$. For the function with the positive slope that crosses the axis, the opposite intransitive cycle occurs. On the other hand, for two functions in the figure that do not cross the horizontal axis, the value of $-a_0/a_1$ is not in the above range. Consequently, for these specifications of the model preferences will be transitive, since osd(d_c) is either always positive, or always negative, in the range $1 < d_c < 4$.

Table 2 illustrates combinations of binary preference in Tversky's lottery paradigm consistent with adherence to, or violation of, transitivity under the algebraic SAD model, for three ranges of $-a_0/a_1$ ($1 < -a_0/a_1 < 2$; $2 < -a_0/a_1 < 3$; and $3 < -a_0/a_1 < 4$), and for positive and negative slopes of the a_1 parameter. These six combinations lead to violations of transitivity for different triples of lotteries, as shown in Table 3. As the table shows, which triples violate transitivity depends on whether $-a_0/a_1$, is positioned between 1 and 2, 2 and 3, or 3 and 4. As discussed later, the same patterns of violation are consistent with parallel specifications of the algebraic difference threshold model.

---Figure 1 in here ---

Table 2

Six combinations of binary preferences for specifications of the algebraic SAD and DT models consistent with violations of transitivity in Tversky's lottery set

Lotteries												
			$1 < t < 2$; $c_1 < 0$ $1 < -a_0/a_1 < 2$; $a_1 < 0$					$1 < t < 2$; $c_1 > 0$ $1 < -a_0/a_1 < 2$; $a_1 > 0$				
P	S	Label	a	b	c	d	e	a	b	c	d	e
7/24	5.00	a	-	1	0	0	0	-	0	1	1	1
8/24	4.75	b		-	1	0	0		-	0	1	1
9/24	4.50	c			-	1	0			-	0	1
10/24	4.25	d				-	1				-	0
11/24	4.00	e										-
			$2 < t < 3$; $c_1 < 0$ $2 < -a_0/a_1 < 3$; $a_1 < 0$					$2 < t < 3$; $c_1 > 0$ $2 < -a_0/a_1 < 3$; $a_1 > 0$				
P	S	Label	a	b	c	d	e	a	b	c	d	e
7/24	5.00	a	-	1	1	0	0	-	0	0	1	1
8/24	4.75	b		-	1	1	0		-	0	0	1
9/24	4.50	c			-	1	1			-	0	0
10/24	4.25	d				-	1				-	0
11/24	4.00	e					-					-
			$3 < t < 4$; $c_1 < 0$ $3 < -a_0/a_1 < 4$; $a_1 < 0$					$3 < t < 4$; $c_1 > 0$ $3 < -a_0/a_1 < 4$; $a_1 > 0$				
P	S	Label	a	b	c	d	e	a	b	c	d	e

7/24	5.00	a	-	1	1	1	0	-	0	0	0	1
8/24	4.75	b		-	1	1	1		-	0	0	0
9/24	4.50	c			-	1	1			-	0	0
10/24	4.25	d				-	1				-	0
11/24	4.00	e					-					-

Notes: 1 = row lottery preferred; 0 = column lottery preferred; a_0 and a_1 are parameters of the SAD model; t and c_1 are parameters of the DT model (see text).

Table 3

Triads of lotteries from Tversky's lottery set for which the algebraic SAD and DT models are consistent with adherence to, or violation of, transitivity for three ranges of the function, -a_0/a_1, of the parameters of the SAD model, and of the threshold parameter, t, of the DT model (V = violate transitivity; otherwise adhere to transitivity)

Lottery Triple	$1 < t, -a_0/a_1 < 2$	$2 < t, -a_0/a_1 < 3$	$3 < t, -a_0/a_1 < 4$
abc	V		
abd		V	
abe			V
acd		V	
ace		V	V
ade			V
bcd	V		
bce		V	
bde		V	
cde	V		
Number of triads predicted to violate transitivity	3	5	3

The stochastic SAD model and probabilistic specifications of transitivity (WST and TI)

We complete the specification of the SAD model by including, as Tversky did for the extended additive difference model, a function that relates the overall subjective difference function of the algebraic model to choice probabilities. We follow one of Tversky's suggestions and specify the logistic function. The stochastic SAD model is satisfied whenever condition (2) and equation (3) hold and

(4) $p(x, y) = \exp(osd(d_c))/(1+ \exp(osd(d_c)))$

where d_c is the objective difference level for lottery pair (x, y), and p(x, y) is the choice probability of choosing x over y.

The SAD model belongs to the broader category of *degree of preference* models, which directly map the strength of preference of one alternative over another to choice probability, without assuming error between preference and choice, or variability of preferences. The SAD model is consistent with violations of WST in the same way that the algebraic SAD model is consistent with intransitive preference cycles; that is, WST is violated when $1 \leq -a_0/a_1 \leq 4$, and conversely, WST is satisfied when $-a_0/a_1$ is outside that range. For example, when $-a_0/a_1 = 1.5$, p(a, b) > .5, p(a, c) >.5 but p(a, c) < .5, which violates WST. Thus, Figure 1 and Tables 2 and 3 also elucidate the relationship between the SAD model and WST. That is, the figure and tables define precisely for which SAD model parameter values (a_0, a_1 and the function $-a_0/a_1$), and for which lottery triples, WST will be adhered to or violated.

Unlike for WST, there is not a straightforward algebraic rule determining when the SAD model is consistent with adherence to, or violation of, TI. Here we develop graphical methods to reveal the precise relationship between the SAD model and both WST and TI. We first present an analysis for each lottery triple, followed by an overall analysis across triples. There are four different dimension difference (d_c) combinations, each with a different set of choice probability triples permitted under the SAD model: (1) abc, bcd and cde, where $d_c = 1$

or 2; (2) abd, bce acd and bde, where d_c = 1, 2 or 3; (3) abe and ade, where d_c = 1, 3 or 4 ; and (4) ace, where d_c = 2 or 4. The graphs in Figure 2 show regions of the parameter space where the SAD model describes adherence to, or violation of, WST and/or TI for each of the ten lottery triples. The figure shows that the graphs for triples within each of the four groups listed above are identical, for example those triples in the first group, abc, bcd and cde. The graphs only show one quadrant of the parameter space, bounded by 0 and 5 for a_0, and by -5 and 0 for a_1, because all the violations described by the model for our five data sets are within this quadrant. The opposite quadrant, bounded by -5 and 0 for a_0, and by 0 and 5 for a_1, also contains regions where violations of the transitivity conditions are described by the model. These are mirror images of the regions in the quadrant shown. In the remaining two quadrants of the parameter space, the model describes adherence to both WST and TI. It can be seen that each graph has a similar structure but different regions of violation; that for WST is within the triangle, and that for TI is within the curve. For each lottery triple there are four regions of the parameter space in which neither, either, or both conditions are violated. Thus, the graphs show how the SAD model can describe violations of both WST and TI, at the level of the lottery triple.

--- Figure 2 in here ---

Turning to the overall analysis, Regenwetter et al. (2010, 2011) explain that both WST and TI are fully specified as conjunctions, across all lottery triples, of the inequality condition that specifies them at the level of the triple. Thus, WST and TI are satisfied if and only if the relevant condition is satisfied for each triple in the set, and, conversely, they are violated if the relevant condition is violated in at least one triple. Figure 3 presents regions of the SAD model parameter space where the model describes overall adherence to, or violation of, TI and WST. The triangle shows the region within which the SAD model parameters describe violation of WST in at least one triple, and the curve shows the region within which TI is violated in at least one triple.

--- Figure 3 in here ---

Nested models

It is useful to define two models that are nested within the SAD model. First, the algebraic constant choice probability (CCP) model is a one-parameter model where the slope of the linear relationship of the SAD model is zero:

(5) $osd(d_c) = a_0$.

In this case the overall subjective difference, or strength of preference, is constant across all ten pairs of the lottery set. Consequently, under the stochastic CCP model, binary choice probabilities are constant across all lottery pairs, i.e. not sensitive to differences between P or S values, and WST will be satisfied. Different values of overall subjective difference, a_0, can be interpreted differently: $a_0 > 0$ reflects a constant strength of preference for the higher S, whereas $a_0 < 0$ reflects a constant strength of preference for a higher P. Additionally, a value of $a_0 \leq -1.39$ represents a very strong preference for the better P and can be interpreted as a stochastic 'take the best P' heuristic. On the other hand, $a_0 \geq 1.39$ represents a very strong preference for the better S, interpretable as a stochastic 'take the best S' heuristic (Gigerenzer & Goldstein, 1999).

The second special case of the SAD model, also nested within the CCP model, is the zero-parameter model in which $a_1 = a_0 = 0$. This is the random choice (RC) model, equivalent to having no preference for one or other of the lotteries of any pair, such that $osd(d_c) = 0$ and $p(x, y) = .5$ across all pairs. In this case, a person is indifferent and chooses at random. The RC model serves an important function in our analysis. As we explain later, it is the first model tested in a series of likelihood ratio (LR) tests of the nested models, i.e., RC, followed by CCP, SAD and the unrestricted model. If RC is a good fit to the data, and no other model is a significantly better fit, we conclude that indifference and random choice is the most parsimonious, well-fitting model for that data (see Ranyard et al., 2020, where such cases were identified). On the other hand, if CCP is a significantly better fitting model, random choice as a mechanism to account for the data can be ruled out, and we can begin to distinguish which of the remaining models, is the better-fitting.

Difference threshold and lexicographic semiorder models

A difference threshold model is a special case of an additive difference model in which the overall subjective difference between two lotteries is a step function. A lexicographic semiorder model is a special case of a difference threshold model which describes intransitive preference cycles.

The algebraic difference threshold model

To specify the algebraic difference threshold model the same simplifications as for the SAD model are applied (see Condition 2). We then specify a step function for the overall subjective difference function between lotteries x and y, $osd_{dt}(d_c)$, with three parameters, t, c_1 and c_2:

(6) If $d_c < t$ then $osd_{dt}(d_c) = c_1$; if $d_c > t$ then $osd_{dt}(d_c) = c_2$,

where d_c is the objective dimension difference level between lotteries x and y; t is the difference threshold parameter in the range $1 < t < 4$; and c_1 and c_2 are in the range -5 to 5.

The function $osd_{dt}(d_c)$ models the difference between the subjective differences of the S and P dimensions in the same way that the $osd(d_c)$ function does for the algebraic SAD model. The difference between them is that the SAD function is linear, whereas the difference threshold model has a step function, as illustrated in Figure 4. Whereas the linear function of the SAD model is compatible with a compensatory process, the step function of the difference threshold model is compatible with a noncompensatory process. As with the SAD model, under the difference threshold model preferences will either be intransitive for some lottery triples, or transitive for all, depending on the parameter values. Specifically, if both c_1 and c_2 are positive, or if they are both negative, preferences will be transitive. However, if $c_1 > 0$ and $c_2 < 0$, or vice versa, preferences will be intransitive for some subset of triples, depending on the value of the threshold parameter t.

We refer to an algebraic difference threshold model with this configuration of parameters as an algebraic lexicographic semiorder model, because preferences switch, at some threshold level of d_c, from the better S to the better P (or vice versa). This is an alternative specification of the lexicographic semiorder model to Tversky's (1969) definition, a specification defined in terms of strength of preference, i.e., overall subjective difference. We define it in this way in order to compare the goodness of fit of the SAD and difference threshold models as explanations of intransitive preferences.

--- Figure 4 in here ---

The relation between the algebraic difference threshold model and transitivity in Tversky's paradigm

The basic similarity between the algebraic SAD and difference threshold models is that they describe when preference adheres to, and when it violates, transitivity in Tversky's paradigm in the same way. Specifically, the lexicographic semiorder variants of the algebraic difference threshold model describe the same intransitive preference cycles for three different ranges of t as does the algebraic SAD model for ranges of the parameter function $-a_0/a_1$. This is illustrated in Tables 2 and 3 in matrix and list representations respectively. Table 2 shows the six combinations of binary preferences which lead to violations of transitivity under different specifications of the lexicographic semiorder model, and Table 3 lists, for the ten triples of lotteries from Tversky's set, the triples which are transitive or intransitive under different specifications of the model. Figure 4 illustrates four specifications of difference threshold models, which are lexicographic semiorder models consistent with different intransitive preference cycles for different values of the model's parameters.

The stochastic difference threshold and stochastic lexicographic semiorder models

To complete the specification of a stochastic difference threshold (SDT) model for Tversky's lottery paradigm, the logistic function is applied to determine choice probability, p(x, y), from the overall difference function, $osd_{dt}(d_c)$, in the same way as it is applied to the $osd(d_c)$ function of the SAD model (equation 4). The SDT model is fully specified, then, by condition 2, and equations 4 and 6. The SDT model and the SAD model make the same predictions of adherence to, or violation of, WST, exactly as stated in Tables 2 and 3. However, the two models can be distinguished in terms of their goodness of fit to the observed choice proportions. They also differ in their relationship with TI, an issue we do not explore here. Cases of the SDT model where the sign of parameters c_1 and c_2 are opposite define the stochastic lexicographic semiorder model. The two models nested within the SAD model are also nested within the SDT model. The CCP model is a one-parameter model where, $c_1 = c_2$ for any value of t, which is fixed, say, at $t = 2.5$. The RC model is nested within the CCP model and has no parameters, since $c_1 = c_2 = 0$ for any t, again fixed at an arbitrary value, e.g. $t = 2.5$.

Method

Frequentist goodness of fit comparison of the SAD and SDT models

In this analysis we extend the analysis of Ranyard et al. (2020) as follows. First, we include a little more detail in our descriptive and initial, basic inferential analysis, for example, the number of observed violations of WST and TI (see the supplemental material). Second, we extend the goodness of fit tests of the SAD model, which applied likelihood ratio (LR) tests to determine the best fitting model out of four nested models (number of parameters in parentheses): M_0 (10), SAD (2), CCP (1) and RC (0). Model M_0 is referred to by Tversky (1969) as the nonrestrictive model, which allows intransitive preferences as it places no restrictions on the ten binary choice probabilities of the lottery set. It therefore has ten parameters, one for each lottery pair. Cavagnaro and Davis-Stober (2014) refer to M_0 as the encompassing, or baseline mode. In this extended analysis we identify when the maximum likelihood estimates (MLEs) of the SAD model's parameters are consistent with adherence to, or violation of, both WST and TI.

The MLEs of each model's parameters are first calculated, along with the goodness of fit statistic, -2lnLL, which is assumed to approximate a chi-square distribution with degrees of freedom equal to the number of parameters of the model. Following this, a series of LR tests of difference in goodness of fit between models is carried out, the test statistic being the difference between the -2lnLLs of the two models in question. This is also assumed to approximate a chi-square distribution, with degrees of freedom equal to the difference in the number of parameters of the models. In comparisons with M_0, if this chi-square value is less than the critical value for $p = .05$ (i.e., $p > .05$), we conclude that the model has a good fit to the data, i.e., the non-restrictive model is not significantly better. In comparisons between the other models, if the chi-square statistic is significant at $p < .05$, we conclude that the model with the higher number of parameters has a significantly better fit. Otherwise, we conclude that the additional parameters do not significantly improve model fit. By the principle of parsimony, we conclude that out of all models with a good fit, the one with the lowest number of parameters is the best fit. Thus, if the CCP model has a good fit, and the SAD model does not have a significantly better fit in comparison with it, we conclude that the CCP model is a good, and the best fit. On the other hand, if the SAD model does have a significantly better fit, we test it against M_0, and if SAD has a good fit (i.e., $p > .05$), we identify the SAD model as a good, and the best fit; otherwise, we conclude that M_0 is the best fitting model. The outcomes of these tests, together with the MLE parameters, determine our categorization of each individual's choice data. When the SAD model is both a good, and the best fit, we categorize it further according to whether the MLE parameters are consistent with adherence to, or violation of, WST and TI.

After applying the above approach to the SAD model, we apply it to assess the goodness of fit of the three-parameter SDT model. We then merge the two analyses and categorize all participants according to whether the best fitting model was the unconstrained model M_0, CCP (including those cases identified as not probabilistic in the exploratory analysis), SAD or SDT. In the case of the latter two, we identify from the MLE parameters whether the model was consistent with adherence to, or violation of either or both of WST and TI. In the final classification we apply the AIC (Akaike information criterion) to test whether the SAD model is a significantly better fit than the SDT model (or vice versa) or whether the data are inconclusive between the two. We adopted a $p < .05$ significance level for these tests.

Finally, only for Tversky's (1969) data, we compare the frequentist analysis of the SAD and SDT models described above with those of the transitive models, WST and TI, derived from the established inequality-constrained methods of Regenwetter et al. (2011) and

Zwilling et al. (2019). While those for TI have previously been published (Regenwetter et al., 2011), the analysis of WST that we present has not.

Pretest predictions of violations of WST

An important test of the validity of a model is to delineate conditions in which participants would be expected to behave in line with the model. In our case it becomes a matter of identifying participants who could be expected to exhibit intransitive preferences in line with the SAD or lexicographic semiorder models. To identify such participants, we follow Tversky's (1969) approach and evaluate the extent to which a certain quantitative indicator derived from the choice proportions in a pretest can predict whether the SAD model consistent with violation of WST is a good fit in other choices from the same participant. We do not estimate the parameters of the SAD model in the first test and check their similarity to those estimated in the second sample, an approach frequently used in model evaluation. Rather, we extend the approach adopted by Tversky, also followed by Montgomery, to two other replications (Regenwetter et al., 2011; Cavagnaro and Davis-Stober, 2014).

As mentioned earlier, Tversky (1969) adopted a two-stage experimental design specifically to make predictions of violation of WST in the second stage. Montgomery (1977) also adopted this design, but the other three studies reviewed here analyzed all of their participants without making such selective predictions. Nevertheless, two of them included additional data sets from the same participants which we can use here to make predictions in the target sets. In the case of Regenwetter et al. (2011), we predict violations of WST in their Cash I data set, which replicated Tversky's lotteries, from choices in their Cash II lottery set in which the lotteries were adjusted to equalize their expected values. In the case of Cavagnaro and Davis-Stober (2014), we predict violations of WST in the data set which replicated Tversky's lotteries in a no time pressure condition, from choices in a separate within-participant time pressure condition using the same lottery set. We can apply a criterion similar to that of Tversky's stage 1 to these additional data sets to predict which participants will violate WST in the target sets. The criterion we used was the following: the participant most often chooses the better S for the smallest dimension difference, i.e., (a, b), (b, c), (c, d) and (d, e), but most often chooses the better P for the largest two-dimension differences, i.e., (a, d), (b, e) and (a, e), or vice versa.

Bayes factor comparison of the SAD and SDT models

Bayes factors provide a method for comparing models which takes into account both the fit of the model, and parsimony, by considering how well the model captures the observed data for all possible parameterizations of the model, weighted by the prior probability of each parameterization (i.e. by considering the marginal likelihood). To calculate Bayes factors for the distributional version of models SAD, SDT and CCP compared to M_0, we estimated the marginal likelihood for each participant for each model using importance sampling (Vandekerckhove, Matzke, & Wagenmakers, 2015). We assumed uninformative uniform priors for all parameters. For continuous parameters the Beta(1,1) distribution was used, with a linear transformation applied for parameters which could take values outside the range of 0 to 1. For the categorical threshold parameter of the SDT model, a categorical prior was used with equal weight on the three possible values. See Table S2 in the supplemental material for the details of the priors and transformations[6]. For the Bayes factors reported in Table S2, we

[6] For both the MLE and BF estimation of the SAD model, both parameters were constrained to be between -5 and 5. This constraint was chosen as: 1) this range covers the qualitatively interesting patterns of behaviour, as seen in the Figures; and 2) by specifying bounds we could place a uniform prior across these values, rather than specifying certain slopes or intercepts as more likely a priori. For consistency across models, the same bounds were used for the c_1 an c_2 parameters in the SDT model, allowing it to predict a similar range of probability strengths. In contrast, rather than parameterise the CCP model as an intercept only specification of SAD, we instead placed a Beta(1, 1) prior directly on the predicted probability, similar to the specification of the 10 parameters of the unconstrained model, M_0. To check the robustness of our results to these assumptions, we conducted three additional model comparisons varying these assumptions (see Table S3 in the supplemental materials). Conclusions were broadly the same across specifications.

used a proposal distribution for each parameter that was a 70:30 mix of the prior and posterior for that parameter and participant[7]. We used rjags (Plummer, 2003) to estimate the posterior distributions running eight chains for 15,000 iterations. For each parameter a beta distribution was fitted to the resulting distribution, so that probabilities could be calculated for the importance sampling procedure.

Bayes factor comparison of transitive and intransitive models

The above Bayesian analyses focus on dimension-based models to clarify the extent of support for the SAD and SDT model consistent with violation of WST and TI, relative to M_0. Next, we present an extended overview of the Bayesian analysis which also includes the transitive models, CCP, WST and TI. Bayes factors for TI compared to M_0 were computed using the order-constrained methods of Heck & Davis-Stober (2019), while those for WST compared to M_0, were computed using the order-constrained methods of Zwilling et al. (2019). The extended overview identifies for each participant which model had the strongest support over all other models listed, and whether the model with the strongest support was strongly supported over all the others (BF > 3). For the SAD and SDT models this is further subdivided according to whether the MLE parameters are consistent with adherence to, or violation of WST and TI. This final analysis in particular identifies the extent of support for SAD or SDT models consistent with violation of WST or TI in comparison to the transitive models.

Graphical analysis of the relationship between the SAD model, WST and TI

For this analysis we plot, for each individual and each lottery triple, the MLEs of the SAD model parameters, a_0 and a_1, onto Figure 2. In each graph of the figure the region of the parameter space within the triangle describes parameter values of the SAD model consistent

[7] Similar results were obtained using mixtures of 0% (i.e. pure prior), 50%, 70% and 90% posterior.

with violation of WST, while that within the curve describes parameter values consistent with violation of TI. The points on the graph represent the SAD model parameter MLEs of those participants from the five data sets (n = 71) for which the parameter values are within the range shown. We also color-code the data points according to whether, for each lottery triple, the observed choice proportions for that triple adhered to, or violate both WST and TI. This gives us a visual representation at the level of the lottery triple of the relationship between violation and adherence to WST and TI under the SAD model and observed violation and adherence to WST and TI.

The remaining participants for whom data is available (n = 19) are not included in the graphical analysis because their MLE parameters are not in the quadrant of the parameter space shown. For these participants, the SAD model MLE parameter values are consistent with adherence to both WST and TI, as were the corresponding observed choice proportions.

Results

Goodness of fit and Bayes factor comparison of the SAD and SDT models

We first report two complementary analyses of Tversky's (1969) data set, beginning with the frequentist approach followed by the Bayesian analysis. We then summarize and compare the analyses of all five data sets. This overview incorporates the analysis explained earlier of predictions of violations of WST for four of the data sets. The detailed analyses of the four replications of Tversky's study are reported in the supplemental material.

Tversky (1969)

Table 4 shows the main outcomes of the frequentist LR analysis for Tversky's (1969) eight participants predicted to violate WST. The columns to the left present the MLE parameter values and goodness of fit statistics for the SDT model. The signs of c_1 and c_2 are opposite for participants one to six, indicating stochastic lexicographic semiorder cases, i.e., describing violation of WST. In contrast, c_1 and c_2 are the same sign for participants 7 and 8, consistent with adherence to WST. The SDT model was a good fit for two of the first six participants, with the asterisks indicating that it was not a good fit for the other four, since the LR test comparing it to M_0 was significant ($p < .05$). The next columns show the equivalent statistics for the SAD model. The LR tests show that this is a good fit, consistent with violations of WST, in six cases. The two columns following show goodness of fit statistics for the CCP and M_0 models, with the former being a good and the best fit for participant 7. In the Category column the classification identifies the best-fitting model, after applying the AIC tests. If these were inconclusive, both models are listed providing they were well-fitting compared to M_0. Finally, the WST and TI columns indicate whether best fitting model's MLE parameters are consistent with adherence to, or violation of WST or TI.

The table shows that for the six participants where the SAD model consistent with violation of WST is a good fit, it is also significantly better than SDT for five of them, as determined by the LR and AIC tests. For three of these participants, the well-fitting SAD model is also consistent with violation of TI. For participant 5, both the SAD and SDT models are a good fit, and the AIC test is inconclusive between them. Finally, the classification shows that models predicting adherence to WST and TI are a good and the best fit in the other two cases, SAD or SDT in one, and CCP in the other.

Table 5 compares the frequentist p-values of goodness of fit of the SAD and SDT models with those of the transitive models, WST and TI, derived from established inequality-constrained methods (Regenwetter et al., 2011; Zwilling et al. 2019). The table shows that WST is a poor fit ($p < .05$) for participants 1, 3 and 6[8], while TI is a poor fit ($p \leq .05$) for participants 3 and 6. The Category column lists all well-fitting models for each participant. This shows that the SAD model consistent with violation of WST and TI is the only well-

[8] This result supersedes the earlier findings of Iverson and Falmagne (1985).

fitting model for participants 3 and 6 (see also Table 1). However, for the other four participants with observed violation of transitive preference (participants 1, 2, 4 and 5) both a model consistent with violation, and one consistent with adherence, are well-fitting models. The Bayesian analysis reported next goes some way towards resolving this.

Table 6 shows the main outcomes of the Bayesian analysis for the same participants; the Bayes factors presented are those for each model compared to M_0. As mentioned, for all data sets, those for the mixture model of transitive preference, i.e., TI compared to M_0, were computed using the inequality-constrained methods of Heck & Davis-Stober (2019). The obtained values closely match those presented by Cavagnaro and Davis-Stober (2014) in their online supplement. Also, for all data sets, Bayes factors for WST compared to M_0, were computed using the order-constrained methods of Zwilling et al. (2019). Again, these correspond very closely to those presented by Cavagnaro and Davis-Stober[9]. Bayes factors for other model comparisons are simply the ratio of those for each model compared to M_0. For example, the Bayes factor for the comparison of WST and TI for participant 8 is 2.69, indicating moderate support for TI over WST, almost reaching the criterion for strong support (BF > 3). Comparing all models with each other in this way enables us to identify the model with the strongest support from the set, shown in the Category column. The final two columns indicate whether the best supported model is consistent with WST and TI respectively. In the case of the SAD model, this is determined by the MLE parameters for that individual.

Table 6 shows that for three participants (1, 3 and 6) the SAD model consistent with violations of both TI and WST has strong Bayesian support relative to the other models, and M_0 has strong Bayesian support when compared to the transitive models, CCP, WST and TI.

[9] We thank Daniel Cavagnaro for providing the BFs for WST not published in the online supplement of Cavagnaro and Davis-Stober (2014), and for guidance on using the QTest 2.1 software.

For participants 2 and 4 the SAD model consistent with violation of WST, but adherence to TI, is supported at least moderately (BF = 1.89 and > 3 respectively), whereas for participant 5 there was strong Bayesian support for the SDT model consistent with violation of WST and adherence to TI. In these latter three cases there was strong Bayesian evidence against WST and CCP in comparison to M_0, (but not TI). Finally, for participants 7 and 8 the model with the strongest support is consistent with adherence to WST and TI; in one case the CCP model is strongly supported, and in the other the SAD model is weakly supported.

Overall, for Tversky's (1969) data the Bayesian results comparing SAD and SDT correspond very closely to those of the LR and AIC frequentist analysis shown in Tables 4 and 5. For three participants (1, 3 and 6) the SAD model consistent with violation of both WST and TI is a good, and the best fitting model, and has strong Bayesian support relative to M_0, while for two participants (2 and 4), the SAD model consistent with violation of WST but adherence to TI is a good, and the best fitting model, and has some Bayesian support relative to M_0. For participant 5, whose choices are consistent with violation of WST (but not TI), the AIC test is inconclusive between SAD and SDT, which are both well-fitting models, while the Bayesian analysis shows strong evidence in support of SDT over SAD. The Bayesian analysis is also more informative since the performance of SAD and SDT can be directly compared with the transitive models, WST and TI. As we saw, the frequentist goodness of fit analyses comparing SAD and SDT with the transitive models finds that both transitive models and those consistent with violation of transitivity are well-fitting models. However, the Bayes factor comparisons show that where the SAD or SDT model consistent with violation of WST or TI is supported, there is also corresponding strong Bayesian evidence against WST or TI. In conclusion, the SAD model provides the better account of intransitive preferences in five cases, and the SDT model does so in one case.

Table 4

Reanalysis of Tversky (1969), r = 20: MLE parameter estimates and goodness of fit (-2lnLL) of the SDT model (left columns), the SAD model (middle columns), CCP and M_0, (right, number of parameters in parentheses)

Part.	t	c_1	c_2	SDT(3)	a_0	a_1	SAD (2)	CCP (1)	M_0(10)	Category	WST	TI
1	2.5	1.06	-.69	***42.81***	2.52	-0.98	***33.73***	72.44	31.75	SAD	No	No
2	2.5	0.41	-.54	47.97*	1.16	-0.52	***44.42***	57.19	33.46	SAD	No	Yes
3	2.5	1.43	-.34	49.64*	2.94	-1.00	***40.65***	78.49	28.38	SAD	No	No
4	2.5	0.23	-2.02	49.97*	1.97	-1.22	***37.98***	87.24	30.77	SAD	No	Yes
5	2.5	0.75	-0.33	***41.18***	1.30	-0.44	***43.87***	53.06	33.27	SAD and SDT	No	Yes
6	3.5	1.22	-1.38	50.56*	2.97	-0.95	***42.84***	76.42	28.13	SAD	No	No
7	1.5	0.25	0.44	42.50	0.20	0.08	42.68	***43.02***	33.93	CCP	Yes	Yes
8	1.5	0.25	1.29	***37.26***	-0.22	0.55	***37.07***	48.19	32.08	SAD and SDT	Yes	Yes

Notes: Part. is participant number; t, c_1 and c_2 are the parameters, and SDT(3) is the goodness of fit (-2lnLL), of the SDT model (3 parameters); a_1 and a_2 are the parameters, and SAD(2) is the goodness of fit (-2lnLL), of the SAD model (2 parameters); * significant departure from M_0, $p < .05$, according to likelihood ratio tests (7 df. for the SDT and 8 df for the SAD model); CCP and M_0 columns show the goodness of fit (-2lnLL) of those models; statistics in bold indicate the model(s) with a good and the best fit; the Category column shows the best-fitting model(s), if both SAD and SDT are a good fit but the AIC test is inconclusive, this is indicated by ‘SAD and SDT’; The WST and TI columns show whether the best fitting model is consistent with adherence o (Yes) or violation of (No) that condition.

Table 5

Reanalysis of Tversky (1969): Frequentist p-values for goodness of fit of SAD, SDT, WST and TI in comparison with the unconstrained model (M_0)

Part.	SAD	SDT	WST	TI	Category	WST	TI
1	.98	.12	.01	.34	SAD, SDT, TI	No	No
2	.20	.04	.10	.59	SAD, WST, TI	No	Yes
3	.14	< .01	.02	.01	SAD	No	No
4	.51	.01	.15	.25	SAD, WST, TI	No	Yes
5	.23	.34	.09	.20	SAD, SDT, WST, TI	No	Yes
6	.07	< .01	.02	.05	SAD	No	No
7	.36	.29	.46	1.00	SAD, SDT, WST, TI	Yes	Yes
8	.50	.64	1.00	1.00	SAD, SDT, WST, TI	Yes	Yes

Notes: Part. is participant number; the Category column shows the models that are a good fit to the data with $p > .05$. The righthand WST and TI columns show whether the best fitting model is consistent with adherence to (Yes) or violation of (No) for that condition.

Table 6

Reanalysis of Tversky (1969): Bayes Factors for comparisons of SAD, SDT, CCP, WST and TI with the unconstrained model (M_0)

Part.	SAD	SDT	CCP	WST	TI	Category	WST	TI
1	1398.26	9.86	> .01	> .01	0.10	SAD	No	No
2	5.21	1.29	0.56	0.03	2.76	SAD (weak)	No	Yes
3	48.15	0.35	> .01	0.02	> .01	SAD	No	No
4	196.35	0.41	> .01	0.27	1.02	SAD	No	Yes
5	6.87	23.02	4.32	0.05	2.10	SDT	No	Yes
6	16.08	0.50	> .01	0.02	> .01	SAD	No	No
7	11.88	27.95	659.81	1.16	16.48	CCP	Yes	Yes
8	262.58	165.43	46.65	6.67	18.24	SAD (weak)	Yes	Yes

Notes: Part. is participant number; the Category column shows the model with strongest support compared to M_0 (BF > 3), if more than one model has strong support compared to M_0, and the BFs do not show one having strong support over the other, the model with the strongest support is listed as having moderate (3 > BF > 2) or weak (2 > BF > 0) support; The righthand WST and TI columns show whether the best fitting model is consistent with adherence to (Yes) or violation of (No) that condition.

Overview of the main Likelihood ratio and Bayesian analyses in relation to WST

Table 7 presents an overview of the LR and Bayesian analyses of SAD, SDT and CCP relative to M_0 for all 116 participants across five studies with respect to adherence to, or violation of WST. Detailed results for each study are given in the supplemental material. The rows of the table labeled P and Not P distinguish, in four of the studies, those participants predicted to violate WST from those not so predicted. The studies by Tversky (1969) and Montgomery (1977) predicted from their pretests that eight and five participants, respectively, would violate WST, with the remaining participants not investigated further (indicated in column Unknown in the table). The three other studies conducted tests of intransitive preferences on the whole sample, without making pretest predictions, but as explained earlier, predictions of violations of WST can be made in the two studies that included additional data from the same individuals (Regenwetter et al., 2011; Cavagnaro & Davis-Stober, 2014).

The LR rows of Table 7 show the number of participants for whom the model indicated in the column was a good, and the best-fitting model, and the BF rows show the number of participants for whom there was Bayes factor evidence in favor of that model. The 'WST violated' columns show the number of cases where only the SAD model or only the SDT model consistent with violation of WST was a good fit, or had strong Bayesian support, or where both were a good fit according to the two sets of likelihood ratio tests, or had strong Bayesian support[10]. The LR rows show the results after the application of the AIC test which identified which of the SDT or SAD model was the better fit, and the BF rows show the number of participants for which there was strong Bayesian support for the SAD model over the SDT model or vice versa. The Both column shows the number of cases that were still

[10] For comparability, we used the MLE parameter estimates in classifying participants in both the likelihood ratio and Bayesian analyses, as these were very close to the modal posterior estimates of the Bayesian analysis.

inconclusive after the AIC and Bayes factor tests. The AIC tests reduced the number of inconclusive cases with respect to the SAD and SDT models from 16 to 9 on the LR analysis, and the Bayesian analysis reduced the number from 15 to 6, showing that the AIC test and the direct Bayes factor comparison between the two models were similarly effective in clarifying cases.

First, we consider the P rows, which tabulate only those participants predicted to violate WST as described earlier. Overall, the P-row subtotals show that a dimension-based model, either SAD or SDT, predicting violations of WST was a good, and the best, fit for 19/29 (65.5%) participants, and similarly, there was strong Bayesian support for such models for 20/29 (69.0%). These proportions were higher for three of the four studies, while for Regenwetter et al.'s study the prediction was fulfilled in only 2/6 cases. The SAD model was rather more successful than the SDT model with respect to both analyses. The Not-P subtotals show that, in contrast, one of the dimension-based models predicting violations of WST was a good, and the best, fit for only 4/31 (12.9%) participants, and similarly, there was strong Bayesian support for only 5/31 (16.7%), excluding the unknown cases.

We now consider the analysis of all 116 participants across five studies. The Total rows show that the unconstrained model, M_0, was the best-fitting model for 18.1% of the sample across all studies according to the LR analysis, whereas according to the Bayesian analysis there was strong support for M_0 for only 10.3%. Turning to the key findings, first consider the LR analysis. The LR totals and percentages rows of Table 7 show that overall, in almost 20% of cases the SAD model describing violations of WST was a good, and the best fit. In under 2% of cases (just two participants) it was the SDT model describing violations of WST that was the best fit, while for about 7% of participants both models were a good fit, with neither being the better fit according to the AIC tests. Overall, in comparison with the unconstrained model, either or both of these two dimension-based models consistent with violations of WST were a good fit for almost 30% of participants over five experiments.

Turning to the Bayesian analysis, the BF totals and percentages of Table 7 show a similar picture. Across all participants of the five data sets there was strong Bayesian evidence supporting a dimension-based model consistent with violation of WST for 35% of participants. The lowest proportion of participants in this category was in Regenwetter et al.'s data set (22%), with the other studies being in the range 24% to 43%. For a substantial majority of these participants the BF evidence supported the SAD model over the SDT model.

It is striking that the findings from both analyses correspond very closely to each other. The relatively minor differences are that the Bayesian analysis found less support for the unconstrained model, and a little more support for both transitive models and dimension-based models consistent with violations of WST.

Table 7

For five experiments, frequencies of participants for whom: (a) the likelihood ratio test (LR) identifies the best-fitting model as the SAD or SDT model expecting violation of WST, or a model expecting adherence to WST (SAD or CCP), or where the unconstrained model is the best fit (M_0); or (b) where there is strong Bayes factor (BF) evidence in favor of the SAD or SDT model consistent with violation of WST, or a model consistence with adherence to WST, or where the unconstrained model is the best fit (M_0); or (c) where the category is unknown because the data is not available

Study	P/Not	LR/BF	WST violated			WST not violated		M_0	Unknown	N
			SAD	SDT	Both	SAD/SDT	CCP			
Tversky (1969)	P	LR	5		1	1	1			8
		BF	5	1		1	1			8
	Not								10	10
Montgomery (1977)	P	LR	2		2			1		5
		BF	2		2			1		5
	Not								16	16
Regenwetter (2011)	P	LR	1		1	2	1	1		6
		BF	2			3	1			6

	Not	LR	1		1	1	7	2		12
		BF	2			1	8	1		12
Cavegnaro et al. (2014)	P	LR	4		3	1		2		10
		BF	5	1	2	1		1		10
	Not	LR	2			3	10	4		19
		BF	1		2	5	10	1		19
Kalenscher et al. (2010	All	LR	8	2	1	3	5	11		30
		BF	14	2		3	6	5		30
Sub-totals	P	LR	12		7	4	2	4		29
		BF	14	2	4	5	2	2		29
	Not	LR	3		1	4	17	6	26	57
		BF	3		2	6	18	2	26	57
Total	All	LR	23	2	9	11	24	21	26	116
		BF	31	4	6	14	26	9	26	116
Percent	All	LR	19.82	1.72	7.76	9.48	20.70	18.10	22.4	
		BF	26.72	3.45	5.17	12.07	22.41	7.76	22.4	

Note: The P and Not rows refer to the participants predicted to violate WST (P), or not so predicted (Not) in the first four studies; the All rows refer to all participants, in Kalenscher et al.'s study because no such prediction was made .

Overview of the LR and BF analyses in relation to TI

Table 8 shows a similar summary of results to the above with respect to the TI condition. As Tversky (1969) made no predictions about violations of this, we present the analysis for all participants. The bottom four rows of Table 8 show the total frequencies and percentages, similar to the previous analyses presented in Table 7 but with respect to TI. We find a substantial minority of participants for whom the best fitting model was a dimensional model (SAD, SDT or both) consistent with violation of TI (15.5%), and for whom there was strong Bayes factor support for such models (23.3%). The aggregate proportions are lower than the corresponding proportions in Table 7 for violation of WST. Nevertheless, this is an important novel finding with respect to Tversky's paradigm, and the concordance between the two analyses is very close. As before, the differences between the LR and Bayesian analyses are relatively minor. Support for M_0 was rather less in the Bayesian analysis, 7.8% compared to 18.1% for the LR analysis, and correspondingly, the percent for which there was strong Bayes factor support was higher for both the SAD model and transitive models. The findings were similar for each data set, except for Regenwetter et al. (2011); in that study transitive models were supported, since there were no participants for whom the SAD model consistent with violation of TI was the best-fitting model, and there was only one instance where there was strong Bayesian support for such a model. In contrast, in the other four studies the percent of participants in this category ranged from 16.7% to 20.7% for the LR analysis, and 16.7% to 31.0% for the Bayesian analysis.

Table 8

For five experiments, frequencies of cases where: (a) the likelihood ratio test (LR) identifies the best-fitting model as either the SAD and/or the SDT model consistent with violation of TI, a model consistent with adherence to TI (SAD, SDT or CCP), or where the unconstrained model is the best fit (M_0); or (b) where there is strong Bayes factor (BF) evidence in favor of the SAD and/or SDT model consistent with violation of TI, or a model consistent with adherence to TI, or the unconstrained model (M_0); or (c) where the category is unknown because the data is not available

Study	Analysis	Violate TI	Adhere to TI		M_0	Unknown	N
		SAD/SDT	SAD/SDT/TI	CCP			
Tversky (1969)	LR	3	4	1	0	10	18
	BF	3	4	1	0	10	18
Montgomery (1977)	LR	4	0	0	1	16	21
	BF	4	0	0	1	16	21
Regenwetter et al. (2011)	LR	0	7	8	3		18
	BF	1	7	9	1		18
Cavagnaro et al. (2014)	LR	6	7	10	6		29
	BF	8	9	10	2		29
Kalenscher et al. (2010)	LR	5	9	5	11		30
	BF	11	8	6	5		30
Total	LR	18	27	24	21	26	116

	BF	27	28	26	9	26	116
Percent	LR	15.52	23.28	20.69	18.10	22.4	
	BF	23.28	24.14	22.41	7.76	22.4	

Bayes factor comparison of the SAD, SDT, CCP, WST and TI models

In the previous presentation of the Bayes factor evidence, we focused on the dimension-based models, including the transitive CCP model, in order to clarify the extent of support for the SAD and SDT models consistent with violation of WST and TI, relative to the unconstrained model, M_0. Here we present an extended overview which also includes the transitive models, WST and TI. The summary is presented in Table 9. The table shows the number of participants for whom each model had the strongest support over all other models listed. In most cases the model with the strongest support was strongly supported over all the others (BF > 3). For the SAD and SDT models this is further subdivided according to whether the MLE parameters are consistent with adherence to, or violation of WST and TI.

The first column of frequencies shows that for all studies, except Regenwetter et al. (2011) for a substantial minority of participants, in aggregate over 20%, the strongest supported model is the SAD or SDT model consistent with violation of both WST and TI. In addition, the second column of frequencies shows that for three studies, including Regenwetter et al., the strongest supported model for over 10% of participants is the SAD or SDT model consistent with violation of WST but adherence to TI. The final three columns show the number of participants for whom a transitive model, consistent with adherence to both WST and TI, was the strongest supported model, or whether M_0 was strongly supported over all other models. An interesting observation here is that in most cases where a transitive model has the strongest, and usually strong (BF > 3) support, the CCP model, and the SAD model consistent with adherence to the WST and TI, were strongly supported over the more general transitive models, WST and TI.

Table 9

Number of participants in each study for whom the model indicated had BF support over all other models. In the case of the SAD and SDT models, the columns show whether the best fitting model parameters were consistent with violation (No WST, No TI) or adherence to (WST, TI) that probabilistic specification of transitivity

Study	SAD/SDT				CCP	WST	TI	M_0	Unknown	N
	No WST		WST							
	No TI	TI	No TI	TI						
Tversky (1969)	3	3		1	1				10	18
Montgomery (1977)	4							1	16	21
Regenwetter et al. (2011)		4	1	2	9		1	1		18
Cavagnaro et al. (2014)	8	3		6	10			2		29
Kalenscher et al. (2010)	11	5		3	6	2	3	0		30
Total	26	15	1	12	26	2	4	4	26	116
Percent	22.4	12.9	0.1	10.34	22.41	1.7	3.4	3.4	22.4	

Note: SAD/SDT indicates either or both of these models; Unknown refers to those participants for whom data is not available; N is the total number of participants in the study.

A graphical analysis of the relationship between the SAD model, WST and TI

Figure 5 shows, for each lottery triple, the regions of the SAD model parameter space within which the parameter values describe violation of WST (the triangles) and TI (the curves). Each point on each graph represents the SAD model parameter MLEs of a participant from the five data sets (n = 71) for which the parameter values are within the range shown. Table 10 summarizes the numerical information representing the correspondence between the SAD model MLEs and the empirical observations shown in Figure 5.

First, consider the points colored in blue which represent individuals for whom the observed choice proportions for that triple violated both TI and WST. For perfect agreement with the SAD model consistent with violations of WST and TI, the blue points should all fall within both the curve and the triangle in each graph. It can be seen in Figure 5 that most of them do fall within (61%), or close to, this region. Second, we can consider the blue and green points together, since in these cases the choice proportions for the triple are observed to violate WST. If these observations are perfectly consistent with the SAD model, they should all fall within the triangles, which again, as seen in the figure, most either do (70%), or are close to it. Third, we can consider the blue and red points together, which are observed to violate TI in the triple concerned. In this case, for perfect agreement with the SAD model these should all fall within the curve of the relevant graph. Again, most do so (69%), although, there are several points outside the curve. Finally, turning to the yellow points, for perfect agreement with the SAD model, these should be outside both the triangle and the curve, as most of them are (84%).

Figure 5 also shows patterns across different triples that are in line with the SAD model. For example, in the graphs for triples abc, bcd, and cde consider the left upper area of the curve which includes violations of both TI and WST. Note that this area is outside the curve and triangle in the graphs for the other triples. The larger area of the curves and triangles for abc, bcd, and cde implies that one should expect more violations of both TI and WST for this group of triples, as compared to the other triples. More specifically, the yellow points (adherence to WST and TI) in the upper left corner of the graphs for the triples with the smaller curves and triangles, for example, abd and ace, in most cases change to blue (violations of both TI and WST) when they are included in the bigger curves and triangles for triples abc, bcd, cde. In general, the points change color in line with the SAD model across different triples when they become covered or not by the curve (violations of TI) and the triangle (violations of WST).

--- Figures 5 in here ----

Table 10

Cross-tabulation of frequencies of lottery triples which according to SAD are expected to be in line with or violate WST or TI (rows) with the corresponding observed frequencies of lottery triples (columns)

	WST TI	WST Not TI	Not WST TI	Not WST Not TI	Sum
WST TI	553	32	11	19	615
WST Not TI	37	55	0	8	100
Not WST TI	56	4	14	10	84
Not WST Not TI	13	23	3	62	101
Sum	659	114	28	99	900

General Discussion

In reviewing previous analyses of choices in Tversky's (1969) lottery paradigm, we found, for a significant minority of participants, strong evidence against models requiring adherence to probabilistic specifications of transitivity, either WST or the mixture model of transitive preference. We concluded from this that violation of transitive preference is a phenomenon that needs to be accounted for. In order to contribute to this, we compared two dimension-based accounts, the SDT model, compatible with a noncompensatory decision process, and the SAD model, compatible with a compensatory process of weighing dimension differences against each other.

Our first step was to extend the SAD model (Ranyard et al., 2020) to determine precisely, by graphical analysis at the level of the lottery triple and overall, the parameter values for which the model is consistent with adherence to, and violation of, WST and TI. The latter is necessary and sufficient for the mixture model of transitive preference for up to five alternatives. Our empirical graphical analysis gave a full and detailed account of the relationship between the observed choice proportions' adherence to, and violation of, WST and TI, and those of the choice probabilities statistically derived from the SAD model. We found that across five studies the observed choice proportions' adherence to, or violation of, both TI and WST were mainly consistent with the SAD model across triples. The novel result from this analysis concerned the relation between the SAD model and TI, which is not straightforward. Specifically, our analysis showed with precision that the dimension-based SAD model accounts well for observed violation of, and adherence to, TI.

We then conducted a detailed reanalysis of Tversky's (1969) lottery study and four incentivized replications recognized as being of high quality. The replications all used Tversky's lottery set adjusted for currency and inflation, and all involved at least ten presentations of each lottery pair. We compared classical frequentist and Bayes factor analyses, and found that they converged on the same key findings. First, with respect to Tversky's study we found that a simplified specification of Tversky's additive difference model consistent with violations of transitivity is a well-fitting model with strong Bayesian support for a significant minority of participants (6/18 for violations of WST, and 3/18 for violations of TI). Furthermore, when we directly compared the Bayes factors of the SAD model consistent with violations of transitivity with those of models consistent with adherence, i.e., WST and TI, we found strong evidence in favor of the SAD model in the same numbers of cases. We also found, using the order-constrained methods of Regenwetter et al. (2011), that WST was a poor fit for 3/18 of Tversky's participants. This previously unpublished result supersedes Iverson and Falmagne's (1985) finding that applied a more conservative test.

Turning to the four replications of Tversky's study, for three of them we find similar strong support from frequentist, Bayesian and graphical analyses for the SAD or SDT model consistent with violations of transitivity. The exception was the Regenwetter et al. (2011) Cash I data set, for which there was strong support for the SAD model consistent with violations of TI for only 1/18 participants. Nonetheless, even in this study, however, we found strong support for the model consistent with violation of WST for 4/18 participants. These findings are novel in two respects. First, the new Bayesian analysis converges with our previously reported frequentist analysis (Ranyard et al., 2020) and shows the extent of support for dimension-based models predicting violations of WST. Second, the graphical, frequentist and Bayesian analyses all show the extent of support for dimension-based models consistent with violations of TI.

In addition to testing how well a model fits the data, testing predictions of a model is important for its evaluation. Tversky (1969) was well aware of this, and his two-stage experiment was designed to test predictions of WST violation from a pretest. Montgomery

(1977) adopted the same approach. We extended this by testing predictions of WST violation in two other replications not specifically designed to do so. For the target data of Regenwetter et al. (2011) and of Cavagnaro and Davis-Stober (2014), we predicted violations of WST from other choices made by the same participants in other data sets. Across the four studies we found that these predictions by a well-fitting dimension-based model, either SAD or SDT, were confirmed for about two-thirds of participants according to both frequentist and Bayesian analyses. In contrast, the proportion was much lower, about 15%, for participants not so predicted in the two studies where the data was available. Although the prediction was not particularly successful in the Regenwetter et al. data set, it was highly successful in the other three studies. In particular, violation of WST in Cavagnaro and Davis-Stober's no time pressure condition was predicted as successfully from choices from the same lottery set in a time pressure condition as was Tversky's prediction from his pretest. Overall, the above-described analyses confirm that the SAD model accounts well for adherence to, and violations of, both WST and TI in Tversky's lottery task.

In accounting for violations of transitivity, our comparison of the performance of a stochastic lexicographic semiorder and a stochastic additive difference model, is arguably an important novel contribution to the literature. Davis-Stober and his colleagues have previously shown that alternative stochastic lexicographic semiorder models are a good fit and better than transitive alternatives for several participants across the experiments reviewed here (e.g., Brown et al., 2015). Our analysis of the SDT model confirms this. It should be noted that the model has relatively loose requirements, i.e., choice probabilities constant in the range .5 to 1.0 below the threshold, constant in the range 0.0 to .5 above the threshold (or vice versa), and with the threshold anywhere in the standard objective dimension difference range of Tversky's lottery paradigm from 1 to 4. This gives the model the best chance of a threshold model fitting the data and, as we have shown, it does so in many cases. In addition, however, we compared its performance to the compensatory SAD model using AIC tests and Bayes factor analysis. These two analyses converge. In particular, with respect to observed violations of WST, the SDT model was the best fit and had strong Bayesian support in comparison with the SAD model for only two participants, whereas the SAD model was the best fit, and had strong Bayesian support compared to SDT, for about 20% of all participants.

As we explained in the introduction, whether a compensatory or a heuristic dimension-based strategy is adopted in Tversky's lottery paradigm has implications for prospect theory. While a lexicographic semiorder account of intransitive preferences is consistent with prospect theory, a compensatory, dimension-based account is not. Our key finding in this regard is that the stochastic lexicographic semiorder model is the best-fitting model for only a small number of participants, while for a large majority of participants, the evidence is consistent with a compensatory, dimension-based evaluation of decision alternatives. This is in contrast to the within-alternative evaluation inherent in both original and cumulative prospect theory. Unfortunately, although the process data of both Montgomery (1977) and Kalenscher et al. (2010) provide evidence of dimension-based processing, they do not discriminate between noncompensatory and compensatory accounts. An exception to this is that Montgomery identified one participant who expressed compensatory evaluation and comparison of within-dimension differences. Notwithstanding this exception, our analysis of the choice data found strong evidence in favor of the compensatory additive difference account for almost all the intransitive participants. In the light of this, previous interpretations of the choice data in Tversky's lottery task, including Tversky's own, should be reconsidered.

Several open questions and avenues for future research remain. First, although dimension-based models were a good fit to over 80 percent of individual choice data, this still leaves over 15% of individuals unaccounted for. Some deviations from the SAD model consist of variations in choice proportion within a given objective dimension difference, which is not permitted under the model. The extent to which such variation can be accounted for by alternative dimension-based models, including ones incorporating error parameters, could be explored. In addition, a more flexible theoretical framework which allows both within-alternative integration and within-dimension comparison of information may be fruitful. A second avenue for future research could apply a combination of process-tracing techniques (Schulte-Mecklenbeck, Kuhberger & Ranyard, 2010). This would contribute more nuanced evidence of the decision processes underlying individual intransitive preferences, particularly whether compensatory or noncompensatory. Thirdly, the issue of changes in decision processes over time should be addressed (Birnbaum, 2022). For example, it is plausible that some participants apply a compensatory decision strategy initially, but drift towards a noncompensatory heuristic as time goes by. Finally, although our findings are limited to a specific simplification of an additive difference model and a specific lottery choice task, they have general implications which need to be investigated further. We do not claim that the SAD model will generalize beyond the specific lottery task for which it was designed. However, Tversky's (1969) additive difference model from which it was derived is a general model, and other variants could be tested in a wider range of tasks where systematic intransitivity has been observed (e.g. Butler & Pogrebna, 2018).

Progress on these theoretical and empirical issues could help us to go beyond our current view, which is as follows. The available decision process and choice data provide convergent evidence that dimension-based decision models account for intransitive preferences in Tversky's lottery task. Furthermore, our classical frequentist and Bayesian analyses of choice data converge to strongly support a compensatory, additive difference account in most instances of intransitive preference, rather than a noncompensatory, dimension difference threshold account.

References

Birnbaum, M. H. & Gutierrez, R. J. (2007). Testing for intransitivity of preferences predicted by a lexicographic semi-order. *Organizational Behavior and Human Decision Processes*, *104(1)*, 96-112.

Brown, N., Davis-Stober, C. P., & Regenwetter, M. (2015). Commentary: "Neural signatures of intransitive preferences". *Frontiers in Human Neuroscience*, *9*.

Budescu, D. V., & Weiss, W. (1987). Reflection of transitive and intransitive preferences: A test of prospect theory. *Organizational Behavior and Human Decision Processes*, *39*(2), 184-202.

Butler, D. J., & Pogrebna, G. (2018). Predictably intransitive preferences. *Judgment and Decision Making*, *13*(3), 217-236.

Cavagnaro, D. R., & Davis-Stober, C. P. (2014). Transitive in our preferences, but transitive in different ways: An analysis of choice variability. *Decision, 1*, 102-122.

Davis-Stober, C. P. (2012). A lexicographic semiorder polytope and probabilistic representations of choice. *Journal of Mathematical Psychology, 56*, 86-94.

Davis-Stober, C. P., Brown, N., & Cavagnaro, D. R. (2015). Individual differences in the algebraic structure of preference. *Journal of Mathematical Psychology,* 66, 70-82.

Edwards, W. (1954). The theory of decision making. *Psychological Bulletin*, *51*, 380-417.

Harte, J. M., Westenberg, M. R., & van Someren, M. (1994). Process models of decision making. *Acta Psychologica*, *87*(2-3), 95-120.

Heck, D. W., & Davis-Stober, C. P. (2019). Multinomial models with linear inequality constraints: Overview and improvements of computational methods for Bayesian inference. *Journal of mathematical psychology*, *91*, 70-87.

Iverson, G., & Falmagne, J. C. (1985). Statistical issues in measurement. *Mathematical Social Sciences*, *10*, 131-153.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk, *Econometrica, 47,* 263-292.

Kalenscher, T., Tobler, P. N., Huijbers, W., Daselaar, S. M., & Pennartz, C. M. (2010). Neural signatures of intransitive preferences. *Frontiers in Human Neuroscience*, *4*.

Lichtenstein, S., & Slovic, P. (2006). The construction of preferences: An overview. In P. Slovice and S. Lichenstein (Eds.) *The construction of preference*, pp. 1-40. Cambridge University Press.

Loomes, G., & Sugden, R. (1995). Incorporating a stochastic element into decision theories. *European Economic Review*, *39*(3-4), 641-648.

May, K. O. (1954). Intransitivity, utility, and the aggregation of preference patterns. *Econometrica: Journal of the Econometric Society* (22), 1-13.

Montgomery, H. (1977). A study of intransitive preferences using a think aloud procedure. In H. Jungermann & G. de Zeeuw (Eds.), *Decision making and change in human affairs*. Dordrecht: Reidel.

Montgomery, H., Selart, M., Gärling, T., & Lindberg, E. (1994). The judgment-choice discrepancy: Noncompatibility or restructuring?. *Journal of Behavioral Decision Making*, *7*(2), 145-155.

Mosteller, F., & Nogee, P. (1951). An experimental measurement of utility. *Journal of Political Economy, 59*, 371– 404.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, *47*(1), 90-100.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), Proceedings of the 3rd international workshop on distributed statistical computing. Vienna, Austria.

Ranyard, R. H. (1977). Risky decisions which violate transitivity and double cancellation. *Acta Psychologica, 41*, 449-459.

Ranyard, R., Montgomery, H., Konstantinidis, E., & Taylor, A. L. (2020). Intransitivity and transitivity of preferences: Dimensional processing in decision making. *Decision*, *7*(4), 287-313.

Ranyard, R., & Svenson, O. (2019). Verbal reports and decision process analysis. In: Schulte-Mecklenbeck, M., Kühberger, A., & Johnson, J. G. (Eds.) *A handbook of process tracing methods*, pp. 270-285. Psychology Press.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2010). Testing transitivity of preferences on two-alternative forced choice data. *Frontiers in Psychology*, *1*, 148.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review, 118*(1), 42-56.

Regenwetter, M., Davis-Stober, C. P., Lim, S. H., Guo, Y., Popova, A., Zwilling, C., ... & Messner, W. (2014). QTest: Quantitative testing of theories of binary choice. *Decision*, *1*(1), 2.

Rieskamp, J. (2008). The probabilistic nature of preferential choice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34(6)*, 1446–1465. https://doi.org/10.1037/a0013646

Savage, L. J. (1954). *The foundations of statistics*. New York: John Wiley.

Slovic, P. (1995). The construction of preference. *American Psychologist*, *50*(5), 364-371.

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review, 76*(1), 31-48.

Tversky, A., & Kahneman, D.(1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty* *5*, 297–323.

Tversky, A., Sattath, S., & Slovic, P. (1988). Contingent weighting in judgment and choice. *Psychological Review*, *95*(3), 371-384.

Vandekerckhove, J., Matzke, D., & Wagenmakers, E. J. (2015). Model comparison and the principle of parsimony. In J. R. Busemeyer, J. Townsend, Z. J. Wang, & A. Aidels (Eds.), Oxford handbook of computational and mathematical psychology (pp. 300–319). Oxford, UK: Oxford University Press.

von Neumann, J & Morgenstern, O. (1944). *The theory of games and economic behavior*. Princeton University Press.

Zwilling, C. E., Cavagnaro, D. R., Regenwetter, M., Lim, S. H., Fields, B., & Zhang, Y. (2019). QTest 2.1: Quantitative testing of theories of binary choice using Bayesian inference. *Journal of Mathematical Psychology*, *91*, 176-194.

Figure captions

Figure 1

The algebraic SAD model: overall subjective difference functions which predict adherence to, and violation of, transitivity: functions a and c predict intransitive cycles for some triples, whereas functions c and d predict adherence to transitivity

Figure 2

Areas of the SAD model parameter space predicting violation of WST and TI in the ranges $-5 < a_1 < 5$, *and* $0 < a_0 < 5$ *for different lottery triples of Tversky's lottery set*

Figure 3

Areas of the SAD model parameter space predicting violation of WST and TI in the ranges $-5 < a_1 < 5$, *and* $0 < a_0 < 5$ *in at least one triple*

Figure 4

The algebraic DT model: overall subjective difference functions which predict adherence to, and violation of, transitivity: functions a and c predict intransitive cycles for some triples (SLS models), whereas functions b and d predict adherence to transitivity

Figure 5

The SAD model MLEs in the ranges $-5 < a_1 < 5$, *and* $0 < a_0 < 5$ *of participants across five data sets, showing, for each triple of Tversky's lottery set, whether the MLEs adhere to, or violate, WST or TI, and whether the observed choice proportions do so*

Figure 1

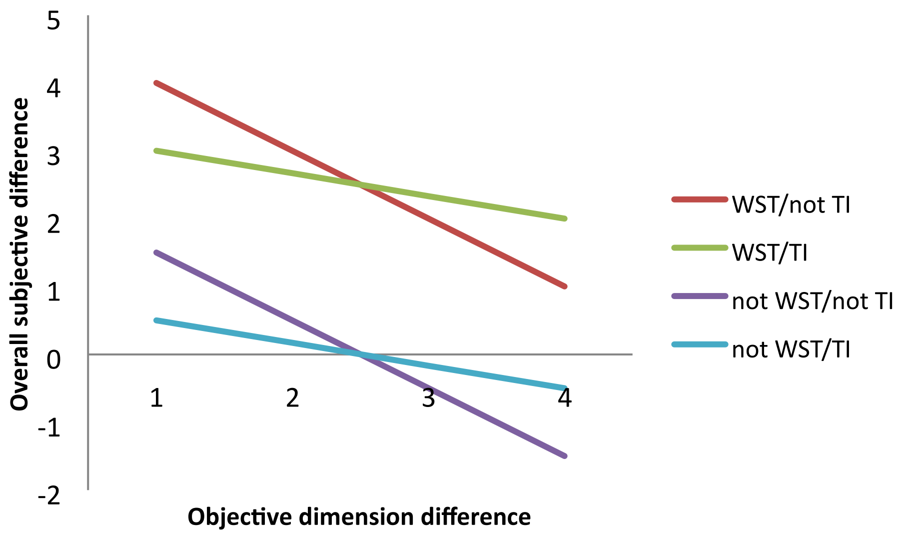

5
4
3
2
1
0
-1
-2
Overall subjective difference
1
2
3
4
Objective dimension difference
WST/not TI
WST/TI
not WST/not TI
not WST/TI

Figure 2

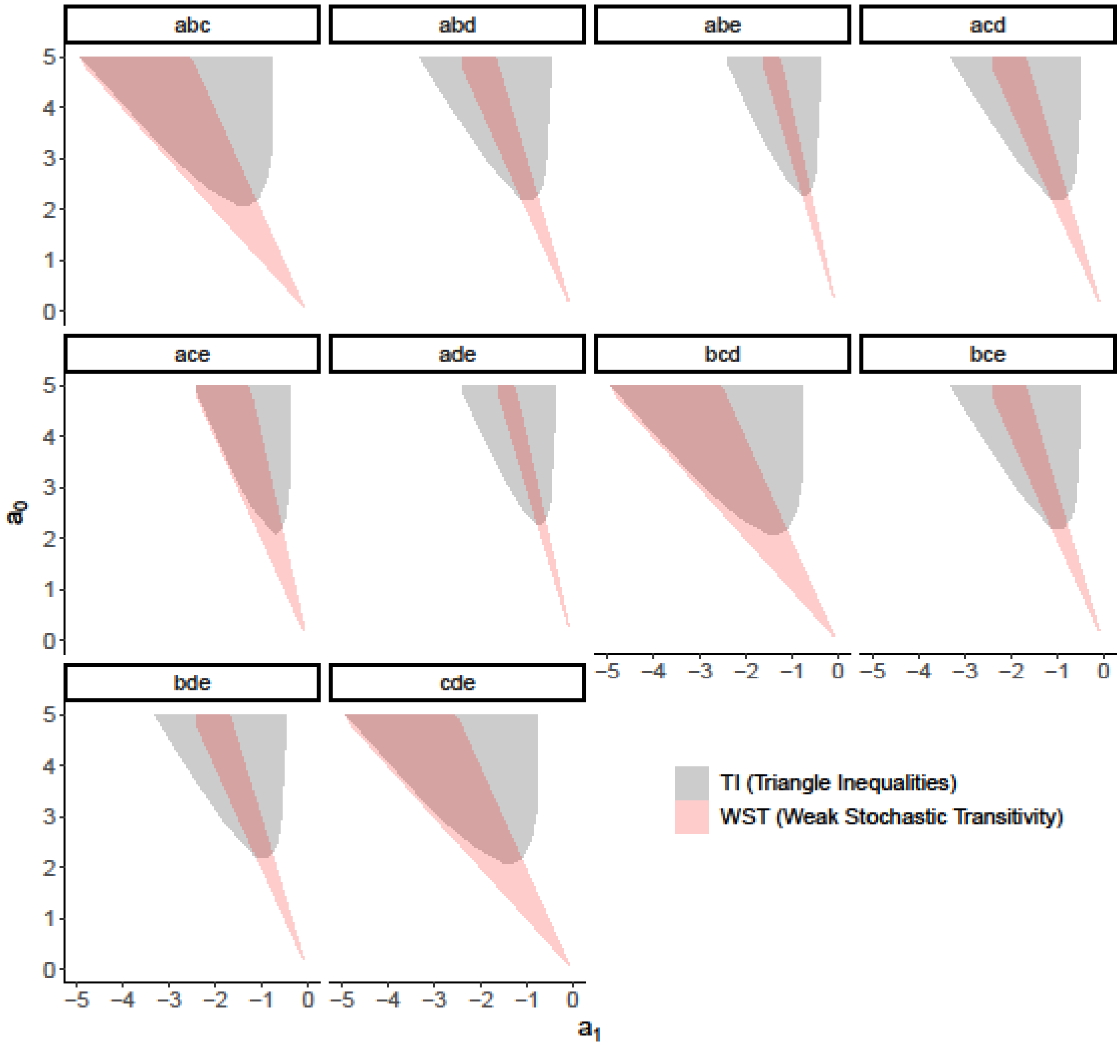

abc
abd
abe
acd
ace
ade
bcd
bce
bde
cde
a0
a1
TI (Triangle Inequalities)
WST (Weak Stochastic Transitivity)

Figure 3

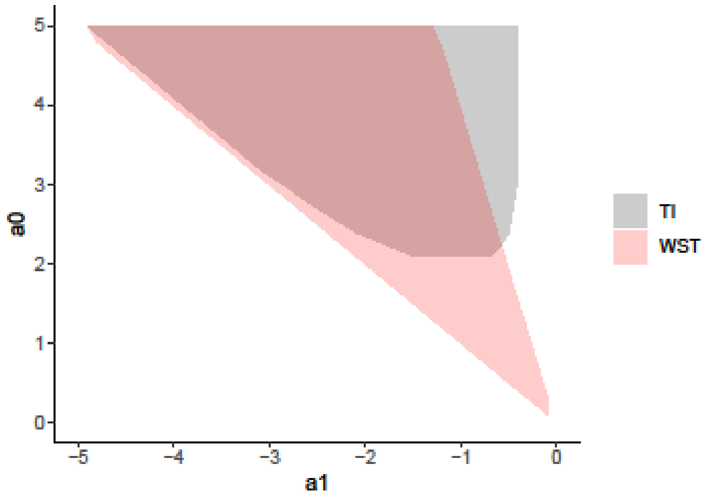


Figure 4

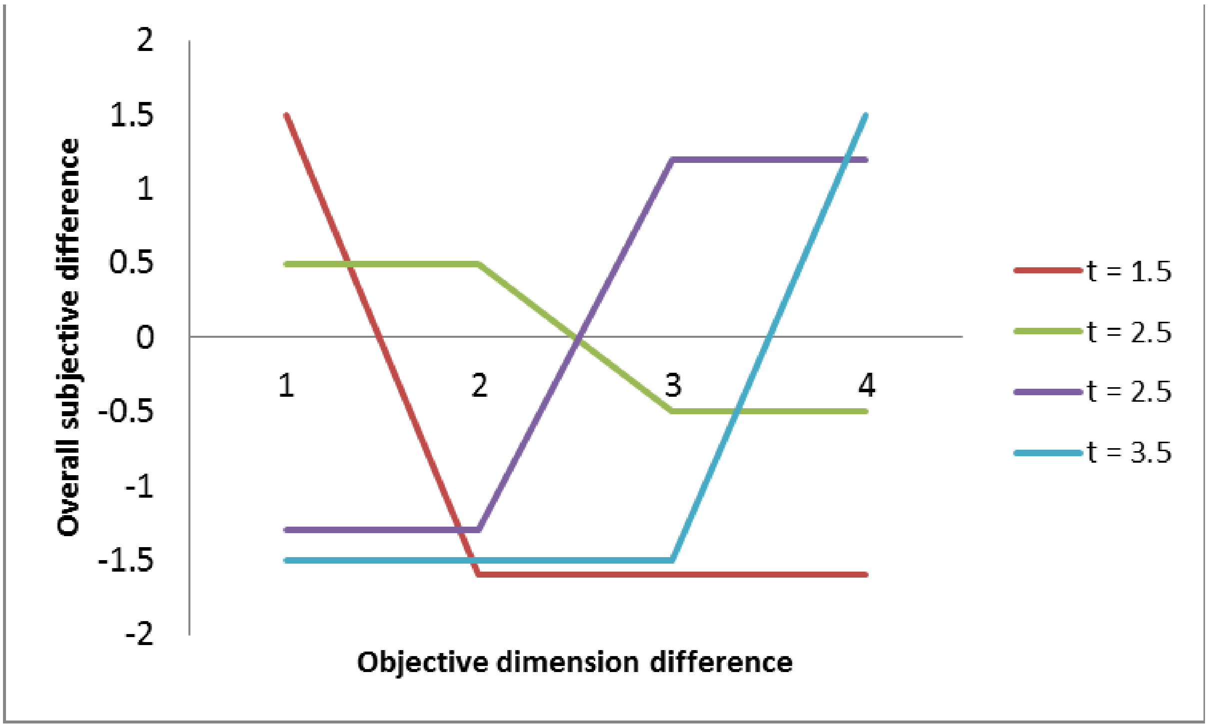
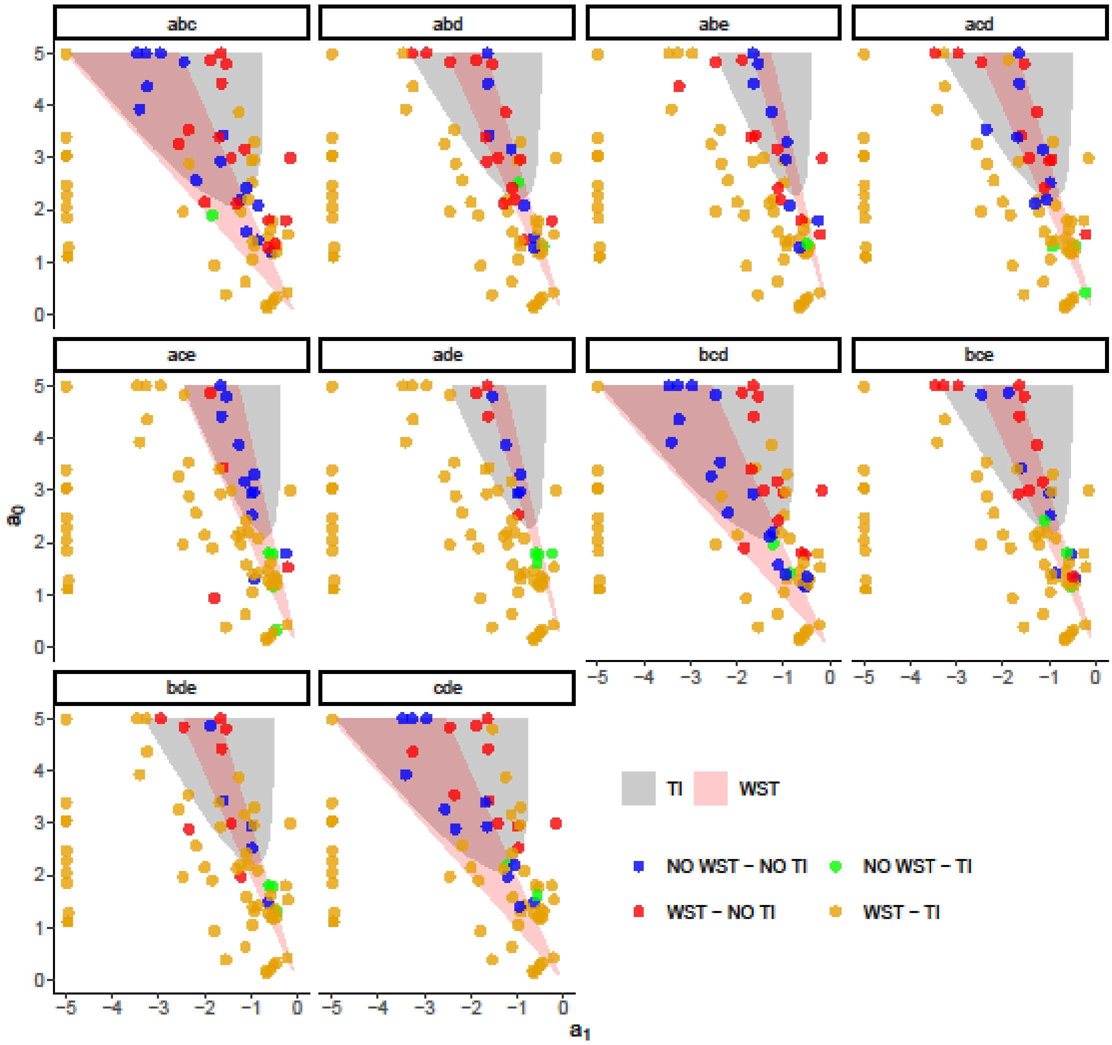

2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
Overall subjective difference
1
2
3
4
Objective dimension difference
t = 1.5
t = 2.5
t = 2.5
t = 3.5

Figure 5


abc
abd
abe
acd
ace
ade
bcd
bce
bde
cde
a_0
a_1
TI
WST
NO WST – NO TI
NO WST – TI
WST – NO TI
WST – TI

Supplemental material: Violations of transitive preference:

A comparison of compensatory and noncompensatory accounts

Rob Ranyard

University of Leeds, UK

Henry Montgomery

Uppsala University, Sweden

Ashley Luckman

University of Exeter, UK

Emmanouil Konstantinidis

University of Warwick, UK

1. Tversky (1969): agreement between SAD model predictions and observed adherence to, and violation of, WST

Table S1 reinforces how well the SAD model accounts for adherence to, and violation of, WST in Tversky's participants one to six. It shows that there is very good agreement between predicted and empirical adherence to, and violations of, WST. Specifically, across these participants, there is agreement in 28 out of 30 combinations of triples and ranges of $-a_0/a_1$. The deviating cases were empirically obtained WST violations for triples ade, bcd, which were predicted to be transitive.

Table S1

Tversky (1969), participants 1-6: number of participants with triples of choice proportions in line with predicted adherence to, or violation of, WST for three ranges of the $-a_0/a_1$, function of the parameters of the SAD model (those not in line with predictions indicated by bold italic numbers).

Lottery Triple	1= < t, $-a_0/a_1$ < 2	2= < t, $-a_0/a_1$ < 3	3 =< t, $-a_0/a_1$ < 4
abc	1	0	0
abd	0	2	0
abe	0	0	1
acd	0	2	0
ace	0	3	1
ade	0	***2***	1
bcd	1	***1***	0
bce	0	4	0
bde	0	3	0
cde	1	0	0
Number of participants	1	4	1

2. Bayes factor prior and posterior distributions

The posterior distributions for all parameters corresponded well with the best fitting parameters obtained from the MLE procedure, as shown in Table S2.

Table S2

Prior distributions used for calculating the BFs in Table 5a. Method correlation and mean difference columns show the correlation, and mean difference, between the mode of the posterior distribution for each parameter and the best fitting MLE parameter, across all participants for the SDT and SAD parameters

Model	Prior	Transformation	Method Correlation	Method Mean Difference
SDT	$t \sim Cat(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	N/A		
	$sc_1 \sim Beta(1,1)$	$c_1 = 10 \cdot sc_1 - 5$	0.997	-0.033
	$sc_2 \sim Beta(1,1)$	$c_2 = 10 \cdot sc_2 - 5$	0.988	0.086
SAD	$sa_0 \sim Beta(1,1)$	$a_0 = 10 \cdot sa_0 - 5$	0.961	-0.146
	$sa_1 \sim Beta(1,1)$	$a_1 = 10 \cdot sa_1 - 5$	0.990	0.049
CCP	$p \sim Beta(1,1)$	N/A		
M_0	$p_i \sim Beta(1,1)$	N/A		

3. Robustness to prior specification

To check how robust our BF conclusions were to our choice of uninformative prior, we conducted 3 additional comparisons. In comparison 2, the SDT model was re-parameterized, so that a Beta(1,1) prior could be placed directly on the two probability parameters, similar to the specification of CCP and M_0. All other models retained their original specification. In comparison 3, the CCP model was re-parameterized as an intercept-only version of the SAD model, with the same uniform prior placed on the a_0 parameter. All other models retained their original specification. Practically, this re-specification places less prior weight on probabilities close to 0.5. In comparison 4, we removed the parameter range constraints in the SAD, SDT and intercept-CCP models, by placing a Normal(0, 2.5) prior on the a_0, a_1, c_1 and c_2 parameters. While this places more weight on values around 0, it keeps >95 percent of the prior mass between -5 and 5, while allowing for more extreme parameter values. Table S3 shows the number of participants, across all datasets, best fit by each model under each of these 4 comparisons, i.e. including the original. Participants are grouped together across studies based on whether they were predicted to: 1) violate transitivity (P); 2) not violate transitivity (Not P); or 3) no predictions could be made in their study (K). While there are some differences, the broad results are consistent across comparisons.

Table S3

Number of participants classified as best fit by SAD, SDT, CCP or M_0, for each of our Bayesian comparisons. Prior specification and/or parameterization of the models differed slightly across comparisons (see text)

P/Not	**Comparison**	**SAD**	**SDT**	**CCP**	**Sat**
K	1- Original	16	3	6	5
	2- Probability SDT	14	7	6	3
	3- intercept-only CCP	16	3	6	5
	4- Normal(0, 2.5) prior	16	3	6	5
Not P	1- Original	8	3	18	2
	2- Probability SDT	5	6	18	2
	3- intercept-only CCP	8	3	18	2
	4- Normal(0, 2.5) prior	9	2	18	2
P	1- Original	20	5	2	2
	2- Probability SDT	15	10	2	2
	3- intercept-only CCP	20	5	2	2
	4- Normal(0, 2.5) prior	20	5	2	2

4. LR test and BF analysis of the four replications

Montgomery (1977)

Table S4 presents the frequentist analysis comparing the goodness of fit of the SAD and SDT models for Montgomery's (1977) five participants who were predicted to violate WST. The columns to the left present the MLE parameter values and goodness of fit statistics for the SDT model. The signs of c_1 and c_2 are opposite for all five participants, consistent with violation of WST. The asterisks indicate that the SDT model was not a good fit for three participants, since the LR test comparing it to M_0 was significant ($p < .05$). The next columns show the equivalent statistics for the SAD model, which LR tests show was a good fit, consistent with violations of WST and TI, in four of the five cases. The penultimate columns show goodness of fit statistics for the CCP and M_0 models. For participant 5, M_0 consistent with violation of both WST and TI, was the best fitting model. The Category column shows the best-fitting model(s). 'SAD and SDT' is recorded in two cases because AIC tests did not identify one as being significantly better than the other. The final two columns show whether the model parameter MLEs were consistent with adherence to, or violation of WST and TI. For all five participants the best fitting model is consistent with violation of both.

Table S5 presents the main findings of the BF analysis of the same data, which shows strong BF evidence supporting dimension-based models consistent with violations of WST and TI for four participants. Specifically, strong evidence in favor of the SAD model over M_0 in all four cases, with strong evidence also in favor of the SDT model over M_0 in three of them. The BFs comparing the two models show strong support for SAD over SDT in three cases ($BF > 3$) and weak support in the other ($2 > BF > 1$). For the remaining participant there is strong evidence in favor of M_0 consistent with violation of both WST and TI. The table also shows strong BF evidence against the three transitive models, CCP, WST and TI, in all five cases. In summary, the BF analysis closely corresponds to the classical frequentist analysis, and in addition clarifies that for all five participants models consistent with violation of WST and TI are strongly supported in comparison with transitive models that require adherence to one or both of these conditions.

Table S4

Reanalysis of Montgomery (1977), r = 10: MLE parameter estimates and goodness of fit (-2lnLL) of the SDT/SLS model (left columns), the SAD model (middle columns), CCP and M_0, (right, number of parameters in parentheses)

Part.	t	c_1	c_2	SDT(3)	a_0	a_1	SAD (2)	CCP (1)	M_0(10)	Category	WST	TI
1	2.5	1.53	-0.69	***26.30***	3.16	-1.13	***25.08***	44.09	21.25	SAD and SDT	No	No
2	2.5	0.83	-1.95	37.82 *	3.40	-1.69	***28.08***	61.69	18.55	SAD	No	No
3	1.5	2.71	-1.10	31.20 *	4.83	-2.46	***23.12***	72.96	14.50	SAD	No	No
4	1.5	2.71	-1.95	***21.17***	5.00	-2.96	***19.75***	80.14	13.21	SAD and SDT	No	No
5	3.5	0.17	-1.95	43.93 *	1.27	-0.64	42.04*	49.50	***21.62***	M_0	No	No

Notes: Part. is participant number; t, c_1 and c_2 are the parameters, and SDT(3) is the goodness of fit (-2lnLL), of the SDT model (3 parameters); a_1 and a_2 are the parameters, and SAD(2) is the goodness of fit (-2lnLL), of the SAD model (2 parameters); * significant departure from M_0, $p < .05$, according to likelihood ratio tests (7 df. for the SLS and 8 df for the SAD model); CCP and M_0 columns show the goodness of fit (-2lnLL) of those models; statistics in bold indicate the model(s) with a good and the best fit; the Category column shows the best-fitting model(s), if both SAD and SDT are a good fit but the AIC test is inconclusive, this is indicated by 'SAD and SDT'; The WST and TI columns show whether the best fitting model is consistent with adherence to (Yes) or violation of (No) that condition.

Table S5

Reanalysis of Montgomery (1977): Bayes Factors for comparisons of models SAD, SDT, CCP, WST and TI with M_0

Participant	M_0	SAD	SDT	CCP	WST	TI	Category	WST	TI
1	1	62.65	22.56	0.12	0.02	0.01	SAD	No	No
2	1	18.02	0.14	1.94×10^{-5}	0.13	0.03	SAD	No	No
3	1	172.46	3.08	6.95×10^{-8}	0.02	0.00	SAD	No	No
4	1	730.31	613.57	1.92×10^{-9}	0.01	0.00	SAD (Weak)	No	No
5	1	0.01	0.01	0.01	0.10	0.07	M_0	No	No

Notes: Part. is participant number; the Category column shows the model with strongest support compared to M_0 (BF > 3), if more than one model has strong support compared to M_0, and the BFs do not show one having strong support over the other, the model with the strongest support is listed as having moderate (BF > 2) or weak (BF > 0) support; The WST and TI columns show whether the best fitting model is consistent with adherence to (Yes) or violation of (No) that condition.

Regenwetter et al. (2011)

Table S6 presents the analysis of goodness of fit of the SAD and SDT models for Regenwetter et al.'s (2011) Cash I lottery set, which directly replicates Tversky's (1969). The participants highlighted in gray (n = 6) are those predicted to violate WST because in set Cash II they switched from majority choice for the better S when dimension difference was smallest, to the better P in the three pairs where it was largest. It can be seen that the prediction of violation of WST was only partially successful. In two of the six cases, as predicted, the SAD model was a good and the best fit, with $1 < -a_o/a_1 < 4$, consistent with violations of WST as described in Table 3. In one of these two cases, the SDT model predicting violation of WST was also a good fit, and the AIC test was inconclusive between these two models. However, for the other four participants models consistent with adherence to WST were a good, and the best fit; the SAD model in two cases, CCP in one, and M_0 in the other.

The non-highlighted cases (n = 12) were not predicted to violate WST on this criterion, and the table shows that models consistent with adherence to WST were supported in most cases. Mean cps (and other criteria) showed that in five cases the better S, and in one case the better P, was chosen almost all the time, which was predictable from the Cash II responses. The other six cases in this subset were subjected to probabilistic analysis. Of these, models consistent with adherence to WST were a good and the best fit for three participants; the CCP model in one case, the SAD model in another, and M_0 in the third. However, models consistent with violations of WST were a good and the best fit for three participants in this subgroup; in one case the SAD model, in another the SAD and SDT models, with the AIC test being inconclusive between them, and in the third case M_0.

Turning to the TI condition, models consistent with adherence to this condition were a good and the best fit for 16/18 participants, seven being the SAD model, eight the CCP model, and one M_0. For the remaining two participants the best fitting model was M_0 consistent with violation of TI.

Table S7 presents the main findings of the BF analysis of the same data. These were in close agreement with those of the frequentist analysis. That is, the prediction of violation of WST was partially successful, with strong evidence supporting the SAD or SDT model predicting violation of WST over M_0 for 2/6 participants. Also, for the cases not predicted to violate WST, only in 2/12 cases was there strong evidence supporting the models consistent with violation of WST. For both of these participants the SAD model had strong support, with strong evidence also supporting the SDT model in one of these. With respect to the TI condition, there was strong BF evidence for the models consistent with adherence to TI for 16 participants, eight being SAD and eight being CCP. Conversely, there was evidence supporting models consistent with violation of TI for two participants.

Overall, then, the two analyses were consistent with each other with respect to support for models consistent with violation of WST. For 4/18 participants the SAD model consistent with violations of WST was a good and the best fit on the frequentist LR analysis, and there was also strong BF support for this model over M_0. With respect to TI, there were only 2/18 participants for whom a model consistent with violation of this was the best model on the LR analysis, M_0 in both cases. The BF analysis was similar; for one participant there was strong support for the SAD and SDT models consistent with violation of TI, and for another there was strong BF support for M_0 consistent with violation of TI.

Table S6

Reanalysis of Regenwetter et al. (2011) Cash I, r = 20: MLE parameter estimates and goodness of fit (-2lnLL) of the SDT model (left columns), the SAD model (middle columns), CCP and M_0, (right, number of parameters in parentheses)

Part.	t	c_1	c_2	SDT(3)	a_0	a_1	SAD (2)	CCP (1)	M_0 (10)	Category	WST	TI
1	1.5	0.10	-1.44	***44.77***	1.05	-0.98	***39.11***	69.03	30.82	SAD (and SDT)	No	Yes
2	1.5	1.24	1.73	43.00*	1.10	0.21	43.54*	44.80	***26.76***	M_0	Yes	Yes
3										CCP	Yes	Yes
4	2.5	0.26	-2.20	87.03 *	2.12	-1.30	74.7*	129.00	***26.85***	M_0	No	No
5										CCP	Yes	Yes
6	1.5	0.05	-2.40	42.84 *	1.90	-1.84	***34.27***	90.25	24.51	SAD	No	Yes
7	3.5	-2.20	-2.94	26.37	-1.90	-0.18	26.44	***26.98***	24.06	CCP	Yes	Yes
8										CCP	Yes	Yes
9	3.5	-0.57	-0.41	46.77	-0.51	-0.02	46.87	***46.89***	33.18	CCP	Yes	Yes
10										CCP	Yes	Yes
11										CCP	Yes	Yes
12	1.5	0.51	-1.10	49.12 *	1.40	-0.96	***44.23***	77.51	31.28	SAD	No	Yes
13	2.5	0.00	-1.49	***45.29***	0.32	-0.47	***46.32***	55.39	32.54	SAD	Yes	Yes
14										CCP	Yes	Yes
15	1.5	0.00	-1.49	46.44*	0.13	-0.65	***35.57***	48.28	30.58	SAD	Yes	Yes
16	2.5	-3.30	-1.39	34.87*	-4.38	0.85	36.23*	48.01	***15.35***	M_0	Yes	No

17	1.5	0.97	0.00	***45.73***	1.35	-0.48	***45.03***	56.01	33.01	SAD and SDT	No	Yes
18	2.5	-.59	-2.20	42.08	0.29	-0.52	***37.90***	48.45	32.42	SAD	Yes	Yes

Notes: see Table S4 for explanation of columns; Gray highlight denotes cases predicted to violate WST from Cash II choice proportions; cases with mean choice proportion > .9 or < .1 classified without further analysis.

Table S7

Reanalysis of Regenwetter et al. (2011), Cash I: Bayes Factors for comparisons of models SAD, SDT, CCP, WST and TI with M_0

Participant	M_0	SAD	SDT	CCP	WST	TI	Category	WST	TI
1	1	113.66	5.69	0.00	3.65	12.73	SAD	No	Yes
2	1	13.25	28.97	212.10	8.43	13.36	CCP	Yes	Yes
3							CCP	Yes	Yes
4	1	$2.19\times^{-6}$	3.35×10^{-9}	1.40×10^{-16}	0.03	0.12	M_0	No	No
5							CCP	Yes	Yes
6	1	2418.09	13.18	3.22×10^{-8}	1.72	5.77	SAD	No	Yes
7	1	119322.4	302768.5	1195029	8.52	8.75	CCP	Yes	Yes
8							CCP	Yes	Yes
9	1	1.52	3.95	93.79	3.10	9.38	CCP	Yes	Yes
10							CCP	Yes	Yes
11							CCP	Yes	Yes
12	1	7.726	0.43	2.12×10^{-5}	0.71	4.17	SAD (Weak)	No	Yes
13	1	2.26	9.92	1.33	2.91	12.96	TI (weak)	Yes	Yes

14							CCP	Yes	Yes
15	1	661.90	463.02	42.02	6.89	18.76	SAD (Weak)	Yes	Yes
16	1	749.08	1603.98	30.88	7.88	0.00	SDT (Weak)	Yes	No
17	1	3.85	2.3	1.00	0.09	1.06	SAD (Weak)	No	Yes
18	1	163.29	90.44	41.82	5.28	19.15	SAD (Weak)	Yes	Yes

Notes: see Table S5 for explanation of columns; Gray highlight denotes cases predicted to violate WST from Cash II choice proportions; cases with mean choice proportion > .9 or < .1 classified without further analysis.

Cavagnaro and Davis-Stober (2014)

In comparison to the above, a similar analysis for Cavagnaro and Davis-Stober's (2014) data set found violation of WST to be rather predictable. In Table S8, the highlighted cases (n = 10) are those predicted to violate WST by choices in the parallel time pressure condition, using the criteria described earlier. The table shows that for most of these (9/10) a model consistent with violations of WST was a good and the best fit, in seven cases the SAD model and in two M_0. The SDT model was also a good fit in six of these cases. The AIC tests comparing the fit of the SAD and SDT models supported the SAD model in three cases but did not differentiate between them in the other three. In the remaining case, the SAD model consistent with adherence to WST was a good and the best fit.

In the subset not predicted to violate WST (n = 19), this was confirmed in most cases. For seven participants the better S, and in one case the better P, was chosen on nearly all occasions, a transitive pattern that was predictable from choices in the time pressure condition. The remaining 12 cases were subjected to probabilistic analysis. Of these, seven participants were best fitted by models consistent with adherence to WST, either the SAD model (three cases), the CCP model (three cases) or M_0. (one case). However, for five participants, models consistent with violations of WST were a good, and the best fit, relative to M_0. For two participants both the SAD and SDT were well-fitting models, with AIC tests finding stronger support for the SAD model in one case, and for three participants M_0 was the best fitting model.

Table S9 presents the results of the BF analysis of the same data, which correspond closely to the above-described LR and AIC test analysis. For seven of the ten participants predicted to violate WST, there is strong BF evidence supporting the SAD and SDT models making that prediction, relative to M_0. For six of these, BFs strongly supported the SAD over the SDT model. Thus, the BFs differentiated between these two models in more cases than did the AIC tests following the LR tests. In addition, there were minor differences between the two analyses for the 19 participants not predicted to violate WST. Whereas in the classical frequentist analysis there were only two participants for whom a dimensional model consistent with violations of WST was a good and the best fit, there was strong BF evidence in support of such models for three participants.

Looking at the sample overall, a dimensional model consistent with violations of WST (either SAD, SDT or both) was a good and the best fit for 31.0% of participants on the classical statistical analysis, with strong BF evidence in favor of such models for 37.9%. In addition, M_0 consistent with violation of WST was the best fitting model for 17.2% of participants and had strong BF support for 6.9%.

Turning to the TI condition, again there were minor differences between the LR and BF analyses. In the former, the SAD or SDT model consistent with violation of TI were a good, and the best fit, for 6/29 (20.7%) of participants, and had strong BF support for 11/29 (37.9%). In addition, M_0 consistent with violation of TI was the best fitting model for 4/29 (13.8%) of participants and had strong BF support in 2/29 (6.9%) of cases.

Table S8

Reanalysis of Cavagnaro & Davis-Stober (2014) Set 1, no time pressure (r = 12): MLE parameter estimates and goodness of fit (-2lnLL) of the SDT model (left columns), the SAD model (middle columns), CCP and M_0, (right, number of parameters in parentheses)

Part.	t	c_1	c_2	SDT(3)	a_0	a_1	SAD (2)	CCP (1)	M_0 (10)	Category	WST	TI
1	3.5	-0.91	-5.00	***33.21***	0.17	-0.67	***32.61***	40.68	24.38	SDT	Yes	Yes
2	1.5	1.47	-0.51	***35.2***	2.42	-1.1	***30.46***	58.77	26.72	SAD	No	No
3	1.5	1.21	-1.95	50.97 *	3.54	-2.36	43.42*	104.81	***16.73***	M_0	No	No
4	1.5	0.00	-3.00	***19.53***	4.99	-5.00	***13.01***	66.08	13.308	SAD	Yes	Yes
5										CCP	Yes	Yes
6										CCP	Yes	Yes
7	1.5	0.00	-3.00	30.2 *	0.93	-1.80	30.03*	48.56	***13.84***	M_0	No	Yes
8	1.5	0.60	-3.00	***20.53***	3.93	-3.41	***24.19***	84.32	14.85	SAD and SDT	No	No
9	1.5	1.34	-0.22	49.03 *	2.08	-0.86	45.24*	63.99	***25.88***	M_0	No	No
10	2.5	-1.16	-0.69	38.16	-1.26	0.16	38.92	***39.30***	25.57	CCP	Yes	Yes
11										CCP	Yes	Yes
12	2.5	0.64	-1.25	***39.81***	2.19	-1.06	***33.24***	59.46	26.65	SAD	No	No
13	2.5	-1.61	-5.00	***22.09***	0.38	-1.55	***19.25***	32.37	15.88	SAD	Yes	Yes

14	2.5	0.05	-2.08	***33.26***	1.30	-0.93	***33.73***	52.42	25.59	SAD and SDT	No	Yes
15	1.5	0.51	-2.40	42.2 *	2.89	-2.34	34.32*	84.00	***16.74***	M_0	No	No
16	3.5	0.61	-1.10	***35.95***	1.61	-0.57	***34.04***	43.04	27.97	SAD and STD	No	Yes
17										CCP	Yes	Yes
18	1.5	1.77	-2.83	29.26 *	5.00	-3.46	***24.16***	117.26	13.05	SAD	No	No
19	1.5	0.42	-2.60	***28.89***	2.57	-2.19	***26.91***	71.19	18.86	SAD	No	No
20	1.5	-1.61	-2.83	31.70 *	-0.71	-0.88	30.17*	35.57	***13.13***	M_0	Yes	Yes
21										CCP	Yes	Yes
22										CCP	Yes	Yes
23										CCP	Yes	Yes
24										CCP	Yes	Yes
25	1.5	0.69	-0.89	***38.13***	1.42	-0.86	***37.20***	54.82	27.00	SAD and SDT	No	Yes
26	1.5	1.77	-0.69	***36.32***	2.99	-1.43	***29.39***	70.31	25.22	SAD	No	No
27	1.5	0.00	-3.00	***24.09***	1.96	-2.46	***17.75***	46.68	16.44	SAD	Yes	Yes
28	3.5	-0.49	-1.61	32.08	-0.22	-0.19	33.54	***34.46***	28.40	CCP	Yes	Yes
29	1.5	1.21	-3.00	34.21*	4.36	-3.25	36.88*	110.31	***13.16***	M_0	No	No

Notes: see Table S4 for explanation of columns; Gray highlight denotes cases predicted to violate WST from the time pressure condition; cases with cp > .9 or < .1 classified without further analysis

Table S9

Reanalysis of Cavagnaro & Davis-Stober (2014) Set 1, no time pressure (r = 12): Bayes Factors (BFs) for SAD, SDT, CCP, WST and TI with M_0

Part.	SDT	SAD	CCP	WST	TI	Category	WST	TI
1	50.05	40.21	20.17	5.8	15.91	SDT (weak)	Yes	Yes
2	6.62	105.35	0.00	0.03	0.20	SAD	No	No
3	0.00	0.34	2.64×10^{-13}	0.07	0.00	M_0	No	No
4	373888	1906215	$5.58\times10\times^{-5}$	4.90	6.32	SAD	Yes	Yes
5						CCP	Yes	Yes
6						CCP	Yes	Yes
7	8730.15	451.93	0.31	7.11	6.02	SDT	Yes	Yes
8	71034.61	8540.66	$6.76\times^{-9}$	1.07	0.23	SDT	No	No
9	0.01	0.06	0.00	0.24	0.32	M_0	No	No
10	3.50	1.31	41.05	3.25	1.00	CCP	Yes	Yes
11						CCP	Yes	Yes
12	0.77	25.64	0.00	0.41	0.68	SAD	No	No
13	15411.74	102979.4	936.97	8.40	17.56	SAD	Yes	Yes
14	21.61	20.23	0.06	1.34	4.02	SDT (weak)	No	No

15	0.26	37.69	8.24×10^{-9}	1.33	1.45	SAD	No	Yes
16	9.36	13.58	6.94	0.50	2.25	SAD (weak)	No	Yes
17						CCP	Yes	Yes
18	279.00	2524.83	5.25×10^{-16}	0.01	0.00	SAD	No	No
19	216.62	1480.27	4.91×10^{-6}	1.69	1.72	SAD	No	No
20	227.79	344.42	178.06	8.12	7.59	SAD (weak)	Yes	Yes
21						CCP	Yes	Yes
22						CCP	Yes	Yes
23						CCP	Yes	Yes
24						CCP	Yes	Yes
25	1.52	3.26	0.02	0.25	1.11	SAD (weak)	No	Yes
26	4.70	218.70	8.41×10^{-6}	0.01	0.02	SAD	No	No
27	22865	250085.1	0.86	7.45	16.20	SAD	Yes	Yes
28	67.90	17.12	496.83	2.52	12.46	CCP	Yes	Yes
29	22.75	10.52	1.66×10^{-14}	0.09	0.00	SDT (weak)	No	No

Notes: see Table S5 for explanation of columns; Gray highlight denotes cases predicted to violate WST from the time pressure condition; cases with cp > .9 or < .1 classified without further analysis

Kalenscher et al. (2010)

Table S10 presents the comparison of goodness of fit of the SAD and SDT models for Kalenscher et al.'s (2010) participants (N = 30). Two participants chose the better S, and two chose the better P, almost all the time. These four were categorized as transitive (CCP model) without subsequent probabilistic analysis. For the remaining 26 participants, the table shows that the SDT model was a good fit in only six cases. For four of these, the signs of c_1 and c_2 were opposite, indicating SLS cases, i.e. the best-fitting model was consistent with violations of WST. The SAD model, on the other hand, was a good, and the best fit in 12 cases, with three being consistent with adherence to WST, and nine consistent with violations. Overall, then, a dimension-based model consistent with violations of WST was a good, and the best fit in 11 cases. AIC tests found that there was strong support for the SAD over the SDT model in 8 of these, with strong support for the SDT over the SAD model in two. In the remaining case both models were a good fit and the AIC test was inconclusive. The unconstrained model, M_0, was the best fit in 11 cases, and overall, transitive models were a good and the best fit in the remaining eight cases, three being the SAD model and five CCP. With respect to the TI condition, a dimensional model consistent with violation of the TI condition was a good and the best fit for 6/30 participants, lower than the 11/30 where this was the case with respect to WST.

Table S11 presents the results of the BF analysis for the same data. The findings are broadly consistent with the LR analysis, though with some differences, mainly because there was strong BF evidence in favor of M_0 for only eight participants, whereas on the LR analysis it was the best fitting model in 11/30 cases. Correspondingly, there was strong evidence in favor of both dimensional models consistent with violations of transitivity for a few more participants, 13/30 (43.3%) with respect to WST, and 6/30 (20.0%) with respect to TI, which is a little more support for these models than in the corresponding LR analysis.

Table S10

Reanalysis of Kalenscher et al. (2010), r = 20: MLE parameter estimates and goodness of fit (-2lnLL) of the SDT model (left columns), the SAD model (middle columns), CCP and $M_{0,}$ (right, number of parameters in parentheses)

Part.	t	c_1	c_2	SDT(3)	a_0	a_1	SAD (2)	CCP (1)	M_0 (10)	Category	WST	TI
1	1.5	1.47	-1.24	51.58 *	2.93	-1.66	44.43*	122.39	***27.22***	M_0	No	No
2	1.5	2.90	-0.10	66.11 *	4.87	-1.89	37.63*	132.12	***20.84***	M_0	No	No
3										CCP	Yes	Yes
4	1.5	2.07	-2.51	36.89 *	5.00	-3.28	***32.17***	185.89	19.19	SAD	No	No
5	2.5	0.99	0.00	48.68 *	1.77	-0.54	***45.48***	58.21	31.98	SAD	No	Yes
6	1.5	0.73	-2.64	46.39 *	3.26	-2.56	40.86*	134.35	***21.27***	M_0	No	No
7	2.5	1.73	-0.77	56.67 *	4.42	-1.64	***31.98***	111.03	25.41	SAD	No	No
8	1.5	-0.85	-.10	49.13 *	-0.67	0.20	57.17*	59.21	***33.18***	M_0	No	No
9	3.5	1.73	-2.20	***32.23***	3.30	-0.97	51.57*	80.32	27.25	SDT	No	No
10	2.5	1.06	-2.02	55.49 *	3.43	-1.61	47.27*	127.48	***26.773***	M_0	No	No
11										CCP	Yes	Yes
12	3.5	1.49	-2.94	44.37 *	3.87	-1.26	43.00*	94.14	***25.94***	M_0	No	No
13	2.5	2.46	-0.34	47.01 *	4.80	-1.54	39.82*	104.14	***22.49***	M_0	No	No
14	1.5	1.03	-0.34	55.84 *	1.46	-0.64	57.79*	76.37	***31.70***	M_0	No	No

15										CCP	Yes	Yes
16	2.5	0.17	-0.48	***47.16***	0.42	-0.22	***49.04***	51.47	33.81	SAD and SDT	No	Yes
17	1.5	1.03	-0.55	***45.28***	1.20	-0.57	57.58*	72.54	32.45	SDT	No	Yes
18	2.5	-1.39	-2.64	***32.62***	-0.54	-0.62	***30.86***	38.99	27.25	SAD	Yes	Yes
19	1.5	1.31	0.00	48.16 *	1.80	-0.62	***45.78***	62.63	31.72	SAD	No	Yes
20	2.5	2.56	-0.41	56.73 *	5.00	-1.65	***32.61***	119.64	18.84	SAD	No	No
21	1.5	0.46	0.00	50.62 *	0.18	0.08	48.57	***48.90***	33.64	CCP	Yes	Yes
22	1.5	0.56	-1.29	***42.73***	1.58	-1.11	***38.80***	79.10	31.34	SAD	No	Yes
23	2.5	0.00	-1.87	51.56*	0.20	-0.57	***41.57***	52.95	31.36	SAD	Yes	Yes
24	3.5	0.45	-1.39	57.68 *	1.21	-0.46	60.50*	70.42	***32.28***	M_0	Yes	Yes
25	1.5	0.97	-0.89	56.22 *	2.19	-1.22	***41.54***	93.52	29.47	SAD	No	No
26	3.5	1.22	0.00	56.16*	1.53	-0.21	58.52*	60.20	***28.00***	M_0	Yes	No
27										CCP	Yes	Yes
28	1.5	0.15	-2.51	43.84 *	2.14	-2.00	***36.53***	98.67	23.82	SAD	No	Yes
29	1.5	0.00	-2.20	***39.52***	0.63	-1.12	***32.78***	58.79	26.84	SAD	Yes	Yes
30	3.5	1.49	-0.20	50.67*	1.79	-0.26	60.09*	62.40	***26.84***	M_0	No	No

Notes: See Table S4 notes for explanation of column contents.

Table S11

Reanalysis of Kalenscher et al. (2010), r = 20:):): Bayes Factors (BFs) for SAD, SDT, CCP and WST with M_0

Part.	SDT	SAD	CCP	WST	TI	Category	WST	TI
1	0.14	10.15	3.94×10^{-15}	0.02	0.00	SAD	No	No
2	0.00	199.04	2.83×10^{-17}	0.00	0.00	SAD	No	No
3						CCP	Yes	Yes
4	428.81	2895.91	6.25×10^{-29}	0.00	0.00	SAD	No	No
5	0.82	3.37	0.32	0.91	2.27	SAD (Weak)	No	Yes
6	2.63	111.58	9.15×10^{-18}	0.16	0.02	SAD	No	No
7	0.013	4364.91	1.05×10^{-12}	0.00	0.00	SAD	No	No
8	0.35	0.01	0.20	0.10	2.14	TI (Moderate)	No	Yes
9	5650.57	0.23	4.46×10^{-6}	0.00	0.00	SDT	No	No
10	0.025	2.30	3.04×10^{-16}	0.00	0.00	SAD (moderate)	No	No
11						CCP	Yes	Yes
12	15.56	18.24	4.70×10^{-9}	0.00	0.00	SAD (weak)	No	No
13	1.83	66.55	3.01×10^{-11}	0.00	0.00	SAD	No	No
14	0.01	0.01	3.84×10^{-5}	0.34	1.64	TI (Weak)	No	Yes
15						CCP	Yes	Yes

16	1.29	0.49	9.85	0.52	8.66	CCP (Weak)	Yes	Yes
17	2.50	0.01	0.00	0.32	1.29	SDT (Moderate)	No	Yes
18	5768.23	10008.54	3711.1	8.48	18.67	SAD (Weak)	Yes	Yes
19	0.78	2.89	0.036	0.45	1.21	SAD (Weak)	No	Yes
20	0.015	947.85	1.29×10^{-14}	1.45	0.00	SAD	No	No
21	8.08	0.62	34.78	1.56	8.95	CCP	Yes	Yes
22	9.64	128.66	9.61×10^{-6}	0.39	5.79	SAD	No	Yes
23	30.37	28.17	4.24	6.84	19.83	SDT (weak)	Yes	Yes
24	0.01	0.00	0.00	1.82	6.41	TI	Yes	Yes
25	0.01	32.72	7.24×10^{-9}	0.48	1.31	SAD	No	No
26	0.03	0.01	0.11	5.29	0.63	WST	Yes	No
27						CCP	Yes	Yes
28	8.38	828.04	4.85×10^{-10}	3.95	1.89	SAD	No	Yes
29	503.56	3858.61	0.21	7.96	19.40	SAD	Yes	Yes
30	0.31	0.00	0.03	2.68	0.02	WST (Moderate)	No	No

Notes: See Table S5 notes for explanation of column contents.

5. Goodness of fit of WST: five data sets

As discussed in the Introduction section, goodness-of-fit tests of WST with more recent inequality-constrained methods are on the published record only for the data of Regenwetter et al. (2010, 2011) In Tables S12-16 we report the frequentist p-values for goodness of fit for five data sets, along with Bayes factors relative to the unconstrained model, and Bayesian p-values. These were computed by the methods of Zwilling et al. (2019) with the QTest 2.1 software, default sampling (Gibbs size 5000, burnsize 1000, Chi-square weights simulation sample size 1000, Random seed 1; in parentheses, random seed 838608).

Table S12

Goodness-of-fit and Bayes factors for WST, Tversky (1969)

Participant	Freq. p	WST:	BF
1	.0054 (.0055)	Very unlikely	.0013 (.0013)
2	.1021 (.0973)	Unlikely	.0317 (.0317)
3	.0172 (.0164)	Very unlikely	.0154 (.0154)
4	.1536 (.1622)	Unlikely	.2599 (.2599)
5	.0903 (.0936)	Unlikely	.0465 (.0465)
6	.0197 (.0200)	Very unlikely	.0198 (.0198)
7	.4698 (.4704)	Very likely	1.9765 (1.9765)
8	1.000 (1.000)	Definitely	6.6653 (6.6653)

Notes:

(1) verbal description of the frequentist p: very unlikely, $p < .05$; unlikely, $.15 > p > .05$; likely/very likely, $p > .15$; definitely, $p = 1$ (that the data could be from the WST model).

Table S13

Goodness-of-fit and Bayes factors for WST, Montgomery (1977)

Participant	Freq. p	WST:	BF
1	.0484	Very unlikely	.0163
2	.3361	Likely	.1343
3	.0314	Very unlikely	.0247
4	.0231	Very unlikely	.0055
5	.3863	Likely	.1052

Table S14

Goodness-of-fit and Bayes factors for WST, Regenwetter et al. (2011), Cash I

Participant	Freq. p	WST	BF
1	1	Definitely	3.6474
2	1	Definitely	8.4313
3	1	Definitely	8.5333
4	.0069	Very unlikely	0.0282
5	1	Definitely	8.5323
6	.6097	Likely	1.7215
7	1	Definitely	8.5248
8	1	Definitely	8.5333
9	1	Definitely	3.1025
10	1	Definitely	8.5322
11	1	Definitely	8.5333
12	.5551	Likely	0.7099
13	.6017	Likely	2.9069
14	1	Definitely	8.5333
15	1	Definitely	6.8948
16	1	Definitely	7.8777
17	.2562	Likely	0.0855
18	1	Definitely	5.2864

Table S15

Goodness-of-fit and Bayes factors for WST, Cavagnaro and Davis-Stober (2014), Set 1, no time pressure

Participant	Freq. p	WST	BF
1	1	Definitely	5.7885
2	.0401	Very unlikely	0.0281
3	.0548	Unlikely	0.0708
4	1	Definitely	4.8984
5			
6			
7	1	Definitely	7.1114
8	.6596	Likely	1.0766
9	.0305	Very unlikely	0.2378
10	1	Definitely	3.2516
11			
12	.2039	Likely	0.4135
13	1	Definitely	8.3992
14	.6661	Likely	1.3431
15	.5106	Likely	1.3003
16	.4586	Likely	0.5016
17			
18	.0745	Unlikely	0.0143
19	.6785	Likely	1.6967
20	.1693	Likely	8.1225
21			
22			
23			

24			
25	.1478	Unlikely	0.2471
26	.0547	Very unlikely	0.0143
27	1	Definitely	7.4541
28	1	Definitely	2.5228
29	.0676	Unlikely	0.0854

Table S16

Goodness-of-fit and Bayes factors for WST, Kalenscher et al. (2010)

Participant	Freq. p	WST:	BF
1	0.0185	Very unlikely	0.0196
2	0.0001	Very unlikely	0.0000
3	1.0000	Certainly	8.5241
4	0.0007	Very unlikely	0.0001
5	0.3956	Likely	0.9141
6	0.2994	Likely	0.1603
7	0.0000	Very unlikely	0.0000
8	0.1575	Likely	0.1041
9	0.0000	Very unlikely	0.0009
10	0.0027	Very unlikely	0.0022
11	1.0000	Certainly	8.5295
12	0.0000	Very unlikely	0.0001
13	0.0002	Very unlikely	0.0005
14	0.5985	Likely	0.3410
15	1.0000	Certainly	8.5270
16	0.4958	Likely	0.5229
17	0.2996	Likely	0.3210
18	1.0000	Certainly	8.4840
19	0.3956	Likely	0.4488
20	0.6606	Likely	1.4513
21	1.0000	Certainly	1.5626
22	0.1467	Unlikely	0.3903
23	1.0000	Certainly	6.8456
24	1.0000	Certainly	1.8234
25	0.2781	Likely	0.4765

26	1.0000	Certainly	5.2870
27	1.0000	Certainly	8.5322
28	1.0000	Certainly	3.9544
29	1.0000	Certainly	7.9644
30	0.2904	Likely	2.6807

5. Descriptive and preliminary inferential statistics

In reviewing these data sets we found it instructive to begin with an exploratory descriptive analysis of each individual set of ten binary choice proportions. We are not alone in this approach; Tukey (1980), for example, advocates the exploration and description of data alongside, or in advance of, advanced, model-based statistical analysis. The descriptive measures we find most useful here are the means and 95% confidence intervals for the ten choice proportions (cps) of the set, and the unstandardized regression coefficient measuring the linear relationship between dimension difference and choice proportion. In addition, we note the cps outside the confidence interval, the probability level of the regression coefficient, and the number of violations of WST and TI. The exploratory analysis identifies cases with minimal choice variability that could be classified on the basis of a simple decision model with a very low error rate; where mean cp is very low or very high, with many cps zero or one, cases are classified as 'take the best S' or 'take the best P'. In these cases, probabilistic model fitting is neither necessary nor appropriate.

Table S17 presents the exploratory analysis of Tversky's (1969) eight participants who undertook the second stage of his lottery study. Mean cp ranged from .43 to .72, indicating that probabilistic modelling is appropriate in all cases. There is a significant linear relationship between dimension difference and cp in seven cases, with only participant eight being a positive relationship. The first six participants exhibit between three and seven violations of TI and WST. Descriptively, then, the choices of the first six participants are in line with expectations from a dimensional model predicting intransitive preferences. Note that this is not merely a subjective impression, it is based on these objective descriptive measures.

The exploratory analyses of three of the other data sets are presented in tables S18, S119 and S20. Some of the main observations from these analyses are as follows.

Montgomery's (1977) five participants (Table S13), and four of Regenwetter et al.'s (2011) 18 participants (Table S14), were in line with expectations from a dimensional model predicting intransitive preferences. Most of the latter's other cases were consistent with a CCP model (e.g., no more than two outlier cps, and cps not sensitive to change in dimension differences), with six choosing either the higher P or the higher S nearly all the time. Cavagnaro and Davis-Stober's (2014) 29 participants (Table S15) fell into three groups: (a) in line with expectations from a dimensional model predicting *intransitive* preferences (12); (b) in line with expectations from a dimensional model describing *transitive* preferences (6); or (c) choice of either the higher P or the higher S nearly all the time (6 participants), or otherwise consistent with a CCP model (5). Since a substantial minority of participants exhibited choices consistent with a dimensional model describing violations of WST, we were motivated to move to the next stage of analysis, evaluation of the goodness of fit of the SAD model.

Table S17

Reanalysis of Tversky (1969): descriptive and preliminary inferential statistics

Part.	mean cp	CI	N out.	B	p	N WST	N TI
1	.62	.40-.80	3	-0.21	.001	5	6
2	.53	.35-.75	5	-0.13	.015	3	1
.3	.69	.50-.85	4	-0.20	.001	4	7
4	.43	.25-.65	5	-0.23	.001	2	2
5	.60	.40-.80	2	-0.10	.029	4	1
6	.72	.55-.90	2	-0.19	.001	3	5
7	.59	.40-.80	0	0.02	.596	1	0
8	.70	.50-.85	1	0.10	.005	0	0

Notes:
(1) Part., participant number
(2) Mean cp, mean choice proportion of the row lottery chosen over the column lottery.
(3) CI, confidence interval, the lower and upper bound of choice proportion for 95% CI for the mean cp.
(4) N outliers, number of choice frequencies outside the CI for the mean cp.
(5) B is the slope of the regression equation, B unit change in cp for 1 unit change in level of difference (if B = .08, for every change difference of 1, choice proportion changes by 7.5%)
(6) p is the probability of observed B under H_0: B = 0.
(7) N WST, number of triads out of 10 violating WST.
(8) N TI, number of triads out of ten violating TI.

Table S18

Reanalysis of Montgomery (1977), r = 10: descriptive and preliminary inferential statistics

Part.	mean cp	CI	N out.	B	p	N WST	N TI
1	.68	.40 - .90	3	-.22	.001	3	7
2	.20	.30 - .80	5	-.30	.001	2	6
.3	.52	.30 - .80	6	-.36	.001	3	8
4	.45	.20 - .70	7	-.37	.001	3	7
5	.48	.20 - .70	3	-.16	.078	2	3

Notes: see Table S1

Table S19

Reanalysis of Regenwetter et al. (2011), r = 20: descriptive and preliminary inferential statistics

Part.	mean cp	CI	N out.	B	p	N WST	N TI
1	.32	.15-.50	5	-.17	.006	0	0
2	.82	.65-.95	2	.01	.291	0	0
.3	.01	.00-.05	0	-.01	.141	0	0
4	.42	.20-.60	8	-.24	.017	2	4
5	.04	.00-.10	0	-.02	.273	0	0
6	.26	.10-.40	5	-.11	.002	1	1
7	.10	.00-.20	0	-.02	.217	0	0
8	.02	.00-.10	1	-.02	.207	0	0
9	.37	.20-.60	2	-.01	.912	0	0
10	.07	.00-.15	0	-.04	.077	0	0
11	.03	.00-.10	1	-.03	.142	0	0
12	.40	.20-.60	4	-.19	.002	2	2
13	.36	.15-.55	0	-.10	.052	1	0
14	.99	.95-1.00	0	.01	.141	0	0

15	.25	.10-.40	2	-.10	.003	0	0
16	.09	.00-.20	2	.07	.066	0	4
17	.59	.40-.80	1	-.11	.018	2	3
18	.34	.15-.55	1	-.10	.004	0	0

Notes: see Table S1; Gray: predicted to violate WST

Table S20

Reanalysis of Cavagnaro and Davis-Stober(2014)): descriptive and preliminary inferential statistics

Part.	mean cp	CI	N out.	B	p	N WST	N TI
1	.26	.08-.50	1	-.11	.013	0	0
2	.55	.33-.83	2	-.33	.001	2	5
.3	.38	.17-.67	8	-.30	.004	2	4
4	.20	.00-.50	3	-.20	.003	0	0
5	.07	.00-.17	1	-.07	.057	0	0
6	.08	.00-.17	0	-.01	.185	0	0
7	.14	.00-.33	2	-.12	.030	0	1
8	.31	.08-.58	8	-.25	.002	3	3
9	.58	.33-.83	3	-.19	.024	3	3
10	.27	.08-.50	0	.03	.549	0	2
11	.03	.00-.08	1	-.00	.154	0	0
12	.52	.25-.75	3	-.23	.001	2	4
13	.12	.00-.25	0	-.09	.001	0	0
14	.39	.17-.67	1	-.18	.002	2	1

15	.30	.08-.58	5	-.25	.009	1	2
16	.63	.42-.83	0	-.10	.008	2	0
17	.05	.00-.17	1	-.05	.128	0	0
18	.37	.17-.58	9	-.34	.001	3	5
19	.28	.08-.50	5	-.23	.001	2	2
20	.10	.00-.25	1	-.06	.118	0	0
21	.02	.00-.08	0	-.02	.402	0	0
22	.05	.00-.17	1	-.05	.056	0	0
23	.01	.00-.08	0	-.01	.402	0	0
24	1.0	.00-.00	0	.00	n/a	0	0
25	.44	.17-.67	3	-.18	.004	3	3
26	.54	.25-.75	6	-.28	.001	0	7
27	.17	.00-.33	2	-.14	.001	0	0
28	.36	.17-.58	0	-.04	.250	0	0
29	.31	.08-.50	6	-.28	.004	2	4

Notes: see Table S1; Gray: predicted to violate WST