Alós-Ferrer, Carlos; Fehr, Ernst; Garagnani, Michele

**Working Paper**

# Identifying nontransitive preferences

Working Paper, No. 415

**Provided in Cooperation with:**
Department of Economics, University of Zurich

*Suggested Citation:* Alós-Ferrer, Carlos; Fehr, Ernst; Garagnani, Michele (2022) : Identifying nontransitive preferences, Working Paper, No. 415, University of Zurich, Department of Economics, Zurich,
https://doi.org/10.5167/uzh-219280

This Version is available at:
https://hdl.handle.net/10419/261391

University of Zurich

Department of Economics

Working Paper Series

ISSN 1664-7041 (print)
ISSN 1664-705X (online)

---

Working Paper No. 415

# Identifying Nontransitive Preferences

Carlos Alós-Ferrer, Ernst Fehr and Michele Garagnani

July 2022

---

# Identifying Nontransitive Preferences

Carlos Alós-Ferrer*[1], Ernst Fehr[1], and Michele Garagnani[1]

[1]Department of Economics, University of Zurich

This version: July 2nd, 2022

**Abstract**

Transitivity is perhaps the most fundamental choice axiom and, therefore, almost all economic models assume that preferences are transitive. The empirical literature has regularly documented violations of transitivity, but these violations pose little problem as long as they are simply a result of somewhat-noisy decision making and not a reflection of the deterministic part of individuals' preferences. However, what if transitivity violations reflect individuals' nontransitive preferences? And how can we separate nontransitive preferences from noise-generated transitivity violations–a problem that so far appears unresolved? Here we tackle these fundamental questions on the basis of a newly developed, non-parametric method which uses response times and choice frequencies to distinguish revealed preferences from noise. We extend the method to allow for nontransitive choices, enabling us to identify the share of weak stochastic transitivity violations that is due to nontransitive preferences. By applying the method to two different datasets, we document that a sizeable proportion of transitivity violations reflect nontransitive preferences. These violations cannot be accounted for by any noise or utility specification within the universe of random utility models. Finally, in spite of revealed transitivity violations, preferences estimated through our method predict choices out of sample better than standard parametric random-utility estimations.

**JEL Classification:** D01 · D81 · D91

**Keywords:** Transitivity · Stochastic choice · Preference Revelation · Predicting Choices

*Corresponding author. Zurich Center for Neuroeconomics, Department of Economics, University of Zurich, Blümlisalpstrasse 10, CH-8006 Zurich, Switzerland. E-mail: carlos.alos-ferrer@econ.uzh.ch

# 1 Introduction

The economic approach to decisions builds upon the assumption that choices can be represented by (complete) transitive binary relations, that is, preferences. Transitivity is hence, arguably, the most fundamental assumption behind economic models of choice. Yet, the empirical literature has regularly documented systematic violations of transitivity in the form of cyclical choices where, for example, $a$ is chosen over $b$, $b$ is chosen over $c$, and $c$ is chosen over $a$ (e.g., Tversky, 1969; Loomes, Starmer, and Sugden, 1989, 1991; Humphrey, 2001).

The interpretation of this empirical evidence is strongly contested. The main argument is that choice is stochastic, and hence it is possible to observe nontransitive choices even though preferences are transitive, because actual choices are noisy (Iverson and Falmagne, 1985; Sopher and Gigliotti, 1993; Birnbaum, 2020). As Birnbaum and Schmidt (2010) observed, *"[a] problem that has frustrated previous research has been the issue of deciding whether an observed pattern represents 'true violations' of transitivity or might be due instead to 'random errors.'"* In other words, while it is tempting to interpret non-transitive choices as evidence of true violations of transitivity (underlying "nontransitive preferences"), those can in principle be explained by, for example, random utility models which postulate a transitive binary relation plus a noise term (McFadden, 1974, 2001; Anderson, Thisse, and De Palma, 1992). The current literature has long been at an *impasse* due to the impossibility of disentangling preferences from noise.

In this contribution, we show how to disentangle preferences and noise to examine whether cyclical choices are due to noise or true evidence of nontransitive preferences. We do this by relying on recent results by Alós-Ferrer, Fehr, and Netzer (2021), which use response times to reveal preferences even when choices alone cannot do so. We extend their results to allow for "preference revelation" even when the underlying binary relation is nontransitive. We then apply the results to two existing datasets (which include both repeated choices and response times) and examine the evidence for violations of transitivity in the light of the new results. In a nutshell, we find that both datasets contain transitivity violations in the underlying preferences, independently of any model of noise. That is, we find a percentage of nontransitive patterns which cannot be explained by any model built upon transitive preferences and noisy choices.

The key to understand our empirical results is the fact that our theoretical approach allows us to examine *Revealed Transtivity Violations* (RTVs) in datasets including repeated choices and response times. Those are patterns of cyclical choices such that, for each choice pair along the cycle, *any* model of preference-

based choice (transitive or not) including noise (no matter which assumptions on the latter are imposed, e.g. symmetric or not), the data reveals that the underlying preference is as specified in the cycle. Hence, the observed cycle can only be explained by a truly nontransitive preference, and not by choice noise. Naturally, not all observed choice cycles are RTVs.

In contrast, the previous literature has concentrated on violations of *Weak Stochastic Transitivity* (WST) in datasets with repeated choices. Denoting by $p(x, y)$ the proportion of $x$ choices from the pair $\{x, y\}$, a WST violation is a pattern in the data where $p(a, b) \geq 1/2$ and $p(b, c) \geq 1/2$, but $p(a, c) < 1/2$.[1] The focus on WST is natural because it is straightforward to show that random utility models, where choices maximize an underlying utility plus a pair-specific noise term, can never violate WST, provided the noise is symmetrically distributed. The latter additional assumption is automatically fulfilled by all standard models used in microeconometric analysis (e.g., probit or logit choice). Hence, we will compare RTVs to violations of WST in both datasets. We show that every RTV implies a violation of WST, but the converse is not true. This is because violations of WST are compatible with *asymmetric* noise and transitive preferences, but RTVs are not.

In order to study transitivity violations, we extend random utility models (RUMs) and their response-times extensions in Alós-Ferrer, Fehr, and Netzer (2021) to allow for nontransitive preferences. This allows to falsify the transitivity hypothesis in models with noisy choices by documenting the existence of nontransitive preferences. To do so, we apply the framework developed in the seminal paper of Shafer (1974), which encompasses models allowing for non-transitive choices such as (generalized) regret theory (Loomes and Sugden, 1982, 1987), salience theory (Bordalo, Gennaioli, and Shleifer, 2012), and Skew-Symmetric-Bilinear utility (Fishburn, 1984a,b,c). The relationship between our approach and previous approaches is as follows. In a standard utility model, $x$ is (weakly) preferred to $y$ if and only if $U(x) - U(y) \geq 0$, where $U$ is a utility function. In a RUM, $x$ is chosen over $y$ if and only if $U(x) - U(y) + \varepsilon_{xy} > 0$, where $\varepsilon_{xy}$ is a pair-specific noise term. In the deterministic model of Shafer (1974), utilities are replaced by two-variable functions $V(x, y)$, which can be thought of as "strength of preference," such that $x$ is (weakly) preferred to $y$ if and only if $V(x, y) \geq 0$. This obviously allows for nontransitive choices, as $V(x, y) > 0$ and $V(y, z) > 0$ do not necessarily imply that $V(x, z) > 0$. In our *Random Nontransitive Models* (RNMs), $x$ is chosen over $y$ if

[1] Note that violations of WST should be tested in experiments or datasets at the individual level, i.e. in settings where the same individual has made a decision multiple times, hence generating choice frequencies.

and only if $V(x,y)+\varepsilon_{xy}>0$, where $\varepsilon_{xy}$ is again a pair-specific noise term. We work in the universe of RNMs and first derive a nontransitive-preference revelation result extending the main result of Alós-Ferrer, Fehr, and Netzer (2021), which we then apply to the data. Models as regret theory or salience theory essentially postulate specific functions $V(x,y)$ capturing certain phenomena (e.g., regret or salience), and thus encompass nontransitive choices. Those models, however, are deterministic, and hence, by definition, cannot tackle noise. Our RNMs encompass all such models while providing a framework where noise can be disentangled from underlying (potentially nontransitive) preferences.

The revelation result we use, as the result of Alós-Ferrer, Fehr, and Netzer (2021), is based on robust empirical regularities of choices and response times arising from psychology and neuroscience. First, easier choice problems are more likely to elicit correct responses than harder problems. This *psychometric effect* is perhaps one of the most robust facts in all of psychology (Cattell, 1893; Laming, 1985; Klein, 2001; Wichmann and Hill, 2001), and extends to cases where the correct response is subjective, e.g. favorite colors, and is uncovered by the researcher through ratings (Dashiell, 1937). The phenomenon has also been established for economic decisions, with evidence dating back to Mosteller and Nogee (1951) and including the recent Alós-Ferrer and Garagnani (2022a,b). Second, easier choice problems take less time to respond to than harder problems. This extremely-robust *chronometric effect* is considered a zero-order fact in the cognitive sciences, and there is overwhelming evidence for it in a wide variety of domains, starting with classical contributions as Cattell (1902), Moyer and Landauer (1967), Moyer and Bayer (1976), and Dehaene, Dupoux, and Mehler (1990). The finding extends of course to preferential choices, as in Dashiell (1937), and a growing number of contributions have demonstrated it in economic decisions, including intertemporal decisions (Chabris et al., 2009), social preferences (Krajbich et al., 2015), and decisions under risk (Moffatt, 2005; Alós-Ferrer and Garagnani, 2022a,b).

Originally, the psychometric and chronometric effects where documented in discrimination tasks, where a decision is hard when the difference between two stimuli is small. The fact that error rates and response times are large in this case simply reflects the difficulty in separating the values of the options (see, e.g., Fudenberg, Strack, and Strzalecki, 2018). In RUMs, harder choices are those where the utilities $U(x)$ and $U(y)$ are closer, and hence more difficult to tell apart. Of course, the psychometric effect is an integral part of standard RUMs, which assume that choice probabilities are monotone in utility differences. The contribution of Alós-Ferrer, Fehr, and Netzer (2021) was to integrate chronometric effects in

RUMs and show how to use them for preference revelation. Analogously, in RNMs, harder choices are those where the strength of preference $V(x, y)$ is smaller, and we rely on psychometric and chronometric effects for our results. Importantly, our approach provides conditions (in terms of choice frequencies and distributions of response times) which, if fulfilled, reveal the underlying preference within a pair independently of any assumptions on the behavioral noise. Within the class of RNMs, those revealed preferences can in turn reveal nontransitive cycles. That is, contrary to WST and other approaches, we do not look for violations of certain implied conditions (on choice frequencies only), but rather examine when nontransitive preferences are revealed by the choice and response time data. In this sense, an RTV does not just imply that the data violates transitivity: it actually reveals nontransitive preferences behind the data.

Our theoretical approach requires datasets where subjects make the same choice multiple times (as in any experiment focusing on WST violations) *and* where response times were explicitly and reliably measured. We obtained two datasets with these characteristics from Davis-Stober, Brown, and Cavagnaro (2015) and Kalenscher et al. (2010). It is important to note that none of these datasets was collected with our approach in mind, and hence they also serve as a demonstration of the applicability of our techniques. As anticipated above, we find that there are revealed transitivity violations in the data, hence rejecting the hypothesis that choices can be represented by transitive preferences plus behavioral noise. Naturally, however, not all violations of WST are true violations of transitivity, and hence our approach provides a better estimate of the extent of nontransitivite preferences, which is necessarily smaller than that derived from WST alone.

The observation that nontransitivities exist and cannot be explained by noise, but that they might be less frequent than previously assumed, begs the question of how much they matter. In the last part of the paper, we use standard microeconometric models *and* the response-times techniques of Alós-Ferrer, Fehr, and Netzer (2021) to predict choices out of sample. Both methods are based on estimating transitive preferences in the presence of behavioral noise, an assumption that is rejected by our analysis. However, the question at this point is not whether this assumption is correct (it is not), but whether it is useful (as a simplification). In other words, we aim to determine whether the fact that transitivity violations exist seriously impairs our capacity to predict new choices. We find that the predictive performance of standard, parametric microeconometric methods is rather modest, but the nonparametric "Time Will Tell" method of Alós-Ferrer, Fehr, and Netzer (2021) significantly improves upon them. The performance is high enough to be

useful, and larger than the standard levels reported in the literature: in spite of nontransitivities, around 75% of out-of-sample choices are correctly predicted by transitive preferences estimated according to the Time Will Tell method.

We view our results as a call for attention. The fundamental assumption that economic choices can be explained by transitive preferences is useful but wrong, even if one allows for behavioral noise. Any model that assumes that people evaluate alternatives independently of other alternatives and tend to choose the option with the higher overall evaluation satisfies transitivity, and hence stands on somewhat-shaky grounds. This includes of course normative models as expected utility theory, but also descriptive models built to accommodate behavioral anomalies as cumulative prospect theory, prospective reference theory, transfer of attention exchange, gains decomposition utility and many others (Tversky and Kahneman, 1992; Birnbaum, Patton, and Lott, 1999; Luce, 2000; Marley and Luce, 2005). The extent of actual transitivity violations in the data might be small enough for those models to remain applicable, but it is clear that their applicability must have an upper bound. Ultimately, applied economics needs to embrace models allowing for violations of transitivity. Those are still sparse (e.g. Shafer, 1974; Loomes and Sugden, 1982; Fishburn, 1982, 1986; Bordalo, Gennaioli, and Shleifer, 2012), but include some prominent examples as salience theory and regret theory.[2] The fact that those models violate transitivity should not be seen as grounds for criticism, but rather as an advantage (a feature, not a bug!).

The paper is structured as follows. Section 2 briefly summarizes the key empirical contributions in the previous literature on transitivity violations. Section 3 presents the theoretical framework, starting with a brief review of the received deterministic models which allow for transitivity violations (Section 3.1) and concluding with our generalization of random utility models to the nontransitive case and the preference revelation result through response times (Section 3.2). Section 4 presents our empirical analysis of two existing lottery-choice datasets and applies the techniques to uncover the extent of revealed transitivity violations. Section 5 presents the out-of-sample prediction analysis. Section 6 concludes. Additional analyses and details are in the (Online) Appendix.

[2]Other models that allow for transitivity violations include lexicographic semiorders (Hausner, 1954; Fishburn, 1971; Birnbaum and Gutierrez, 2007), similarity theory (Fishburn, 1991; Leland, 1994, 1998), the context-dependent model of the gambling effect (Bleichrodt and Schmidt, 2002), and the stochastic difference model of González-Vallejo (2002).

# 2 Previous Evidence on Nontransitivities

Systematic empirical evidence on transitivity violations goes back to May (1954), who collected choice data for pairs of hypothetical marriage partners described according to intelligence, looks, and wealth. However, the evidence was in the form of intransitive cycles when the choices of all participants were aggregated, and hence reduces to the well-known observation that Condorcet cycles might appear when transitive preferences are aggregated. Actual evidence on nontransitive preferences at the individual level was first presented by Tversky (1969), using binary choices among simple monetary lotteries and also among hypothetical job applicants. Almost all participants displayed at least one weak stochastic transitivity violation. These descriptive findings were subsequently replicated (Montgomery, 1977; Lindman and Lyons, 1978; Budescu and Weiss, 1987), but the later literature cast doubts on the strength of the evidence. Iverson and Falmagne (1985) reanalyzed the data of Tversky (1969) and argued that the evidence was compatible with transitive preferences and noisy choices. They further criticized the original work's statistical analysis and found that only one of Tversky's participants significantly violated transitivity using likelihood ratio tests, which of course implicitly assume (a particular shape of) noise in actual choices. It has also been criticized that participants in Tversky (1969) were pre-selected.

Later empirical demonstrations of nontransitive choice have been similarly criticized, the core argument frequently being that data might be compatible with transitive but noisy behavior. For example, Loomes, Starmer, and Sugden (1989, 1991) argued that the classical preference reversal phenomenon (Lichtenstein and Slovic, 1971; Grether and Plott, 1979; Tversky and Thaler, 1990), where choices systematically contradict elicited (monetary) valuations, might be due to transitivity violations. That is, actual nontransitive choices might build a cycle where a lottery $A$ is preferred to a lottery $B$ and this second lottery is (of course) revealed indifferent to its own certainty equivalent, but the latter is strictly preferred to the certainty equivalent of $A$. However, Sopher and Gigliotti (1993), in a replication of Loomes, Starmer, and Sugden (1991), estimated an econometric model of choice with a specific structure of random errors, and could not reject the null hypothesis of transitive preferencess and noisy choices. On the other hand, Starmer and Sugden (1998) further replicated the work in Loomes, Starmer, and Sugden (1991) and observed the same cycling asymmetries, suggesting that those are unlikely to be due to noise. Other arguments which might explain transitivity violations in decisions under risk were considered by Humphrey (2001), who for instance discarded that those might be explained by event-splitting effects (a phenomenon

where preferences are affected by presenting the same event as two different events with the same consequences and the same total probability).

Regenwetter, Dana, and Davis-Stober (2010, 2011) argued that violations of transitivity are better analyzed through violations of the triangle inequality, $p(x, y)+ p(y, z) - p(x, z) \leq 1$ (Marschak, 1960; Block and Marschak, 1960), instead of violations of Weak Stochastic Transitivity. Those works found that the first criterion is often satisfied in (many) existing publications, even when WST is violated. Cavagnaro and Davis-Stober (2014) argued that the tested populations are best described as a mixture of different models of choice, with the resulting estimates suggesting that the majority (but not all) of the people might satisfy transitivity.

Recent studies, however, keep bringing up empirical evidence which might indicate violations of transitivity. Butler and Pogrebna (2018) provided new empirical evidence using both WST and the triangle inequality. Their evidence showed that cycles can be the modal preference patterns over simple lotteries even after considering transitive, stochastic models. Their choices were designed to reproduce the "paradox of nontransitive dice," where a heuristic which favors the option (within a pair) with the largest probability to beat the alternative produce cyclical choices (Savage Jr., 1994). As in previous cases, however, critical work was close on the heels of Butler and Pogrebna (2018). Specifically, Birnbaum (2020) argued that tests of weak stochastic transitivity and the triangle inequality do not provide a method to compare transitive and nontransitive models that allow mixtures of preference patterns and random errors. Birnbaum (2020) re-analyzed the data of Butler and Pogrebna (2018) using a "true and error" model (a class of choice models with noise terms; e.g., Birnbaum, 2013; Birnbaum and Wan, 2020) and still found evidence for significant transitivity violations, but the latter are incompatible with the explanation proposed by Butler and Pogrebna (2018) (see, however, Butler, 2020).

Observed violations of transitivity, whatever their origin, seem to be relatively stable. For example, Davis-Stober et al. (2019) and Park et al. (2019) report that neither age nor, surprisingly, alcohol intoxication seem to play a major role in transitivity violations for decisions under risk. Non-transitive choices have also been observed in other domains. Li and Loomes (2022) report a substantial level of nontransitive choices in respondents' intertemporal decisions, i.e. decisions between pairs of monetary amounts to be received at different points in time (see also Tversky, Slovic, and Kahneman, 1990). Birnbaum and Schmidt (2008) find some evidence for transitivity violations for choices under uncertainty, albeit for a limited number of participants. Moreover, people frequently violate transitiv-

ity when choosing between multi-attribute consumers' products (sound systems, flight plans, and software packages; e.g. Lee, Amir, and Ariely, 2009; Müller-Trede, Sher, and McKenzie, 2015; Lee et al., 2015). Naturally, there are also some domains where evidence is less robust, e.g. for hypothetical alternative treatments in the health domain (Schmidt and Stolpe, 2011), or when choosing between potential sexual partners (Hatz et al., 2020). Finally, violations of transitivity are no exception to the rule that few behaviors, if at all, are uniquely human: honey bees and gray jays have been shown to violate transitivity when foraging for food (Shafir, 1994; Waite, 2001).

We remark that, in this work, we follow the literature which favors testing transitivity violations using binary choice probabilities instead of choice patterns (e.g., Birnbaum, 2020; Birnbaum and Wan, 2020). For a discussion of these two alternative approaches, we refer the reader to Cavagnaro and Davis-Stober (2014) and Butler (2020). This is a natural choice given our theoretical framework, which reveals preferences using binary choices. Moreover, the two approaches have been shown to provide largely consistent evidence (e.g., Butler and Pogrebna, 2018; Birnbaum, 2020).

Needless to say, this section is not and cannot be a complete review of the literature on transitivity violations. We refer the reader to the recent review of Ranyard et al. (2020), who also estimated a simplified additive-difference model based on the processing of alternative dimensions. Similarly to Regenwetter, Dana, and Davis-Stober (2010, 2011), Ranyard et al. (2020) argue that people seem to behave according to different models of choice, and many individuals are best explained by models which do violate transitivity.

# 3 Distinguishing Noise from Nontransitive Preferences

To test whether choices are transitive, one needs to allow for the possibility that they are not. Following Shafer (1974) and others, we refer to a complete but not necessarily transitive binary relation on a set $X$ as a *nontransitive preference*. In this section we first briefly review deterministic models of nontransitive choice, and then proceed to extend random utility models to allow distinguishing between the nontransitivities which are simply due to noise and those which are due to underlying nontransitive preferences.

## 3.1 Deterministic Models of Nontransitive Preferences

If transitivity does not hold, choices can not be presented by preferences or utility functions. It is, however, possible to represent nontransitive binary relations on a set $X$ through real-valued, two-argument functions as follows. Consider a skew-symmetric function $v : X^2 \mapsto \mathbb{R}$, i.e. $v(x, y) = -v(y, x)$ for all $x, y \in X$. We say that a nontransitive preference $\succeq$ on $X$ is represented by a function $v : X^2 \mapsto \mathbb{R}$ if, for all $x, y, \in X$, $v(x, y) \geq 0$ holds if and only if $x \succeq y$. For Euclidean spaces, Shafer (1974) proved that every strictly convex and continuous nontransitive preference can be represented by a continuous, skew-symmetric function as above. This is a natural generalization of representation results for transitive preferences, in which case one can set $v(x, y) = u(x) - u(y)$ for a utility function $u$. Interestingly, the function $v$ has been interpreted as a "strength of preference" (see, e.g. Fishburn, 1988, Chapter 3.9 and ff.), with values of $v(x, y)$ close to zero indicating a difficult decision (the decision maker is close to indifference).

The reason why this representation allows for nontransitivities is transparent. That $v(x, y) \geq 0$ and $v(y, z) \geq 0$ delivers no implication for the sign of $v(x, z)$, while $u(x) - u(y) \geq 0$ and $u(y) - u(z) \geq 0$ immediately yield that $u(x) - u(z) = [u(x) - u(y)] + [u(y) - u(z)] \geq 0$.

When the alternatives are lotteries, adding the requirement that $v$ is linear in both arguments results in skew-symmetric bilinear (SSB) representations, which have been studied by Kreweras (1961) and Fishburn (1982, 1984b, 1986), among others. Specifically, let $L_1, L_2$ be simple lotteries on the set of outcomes $X$, i.e. $L_1(x), L_2(x)$ denote the respective probabilities of outcome $x$ and those are only positive for finitely many outcomes. The function $v$ can be extended bilinearly to simple lotteries by

$$V^{SSB}(L_1, L_2) = \sum_{x \in X} \sum_{y \in X} L_1(x) L_2(y) \cdot v(x, y).$$

so that $L_1$ is weakly preferred to $L_2$ if and only $V^{SSB}(L_1, L_2) \geq 0$. This generalizes expected utility, since if $v(x, y) = u(x) - u(y)$ for a utility function $u$ on $X$, then $V^{SSB}(L_1, L_2) = \sum_{x \in X} L_1(x)u(x) - \sum_{y \in X} L_2(y)u(y)$. However, the SSB form does not require transitivity and indeed allows for cycles and violations of the independence axiom (see Fishburn, 1988 for an axiomatization of SSB nontransitive preferences). That is, the function $V^{SSB}$ is a particular example of the approach of Shafer (1974) for a space of lotteries.

Some other prominent theories have incorporated behavioral phenomena (regret and salience, respectively) in decision making under risk by capturing said

phenomena in a skew-symmetric function over outcomes and then extending it to lotteries in a manner akin to SSB models. Those models, however, are formulated in terms of *acts* (Savage, 1954), that is, mappings from a set of states to outcomes, and hence it is better to change notation at this point. Let the (finite) set of states be denoted by $S$, and let $p(s)$ denote the probability of a state $s \in S$. A lottery $L^x$ is then a vector of outcomes $(x_s)_{s \in S}$, with the interpretation that outcome $x_s$ obtains if state $s$ occurs.

Loomes and Sugden (1982) introduced *regret theory* as a particular model allowing for transitivity violations in the risk domain. Diecidue and Somasundaram (2017) showed that regret theory deviates from expected utility only by relaxing transitivity. Loomes and Sugden (1987) later extended this framework to *generalized regret theory*. This theory considers monetary consequences, $X \subseteq \mathbb{R}$, and starts out by postulating a real-valued, two-argument function $M$, so that if $x, y \in X$, $M(x, y)$ is interpreted as the utility of choosing $x$ net of the regret associated with missing out on $y$. Then it defines the function $v^R$ by $v^R(x, y) = M(x, y) - M(y, x)$ which is immediately skew-symmetric and hence a particular case of the approach of Shafer (1974) for the space of outcomes. Analogously to SSB models, but within the formalization of lotteries as acts, a lottery $L^x$ is weakly preferred to a lottery $L^y$ if and only if $V^R(L^x, L^y) \geq 0$, where

$$V^R(L^x, L^y) = \sum_{s \in S} p(s) v^R(x_s, y_s).$$

Loomes and Sugden (1987) further impose several assumptions on $v^R$, namely that $v^R(x, y) \geq 0$ if and only if $x \geq y$ (so that $v^R$ represents the preferences on outcomes "more is better" in the sense of Shafer, 1974), that $v^R(x, z) > v^R(y, z)$ (resp. $<, =$) if and only if $x > y$ (resp. $<, =$), and a "regret aversion" assumption stating that $v^R(x, z) > v^R(x, y) + v^R(y, z)$ whenever $x > y > z$, meaning that large post-decision regrets are worse than the sum of step-wise, smaller regrets. In particular, skew symmetry and these conditions imply that $v(x, y) > 0$ if $x > y$, $v(x, y) < 0$ if $x < y$, and $v(x, x) = 0$, for any outcomes $x, y$.

The comparison of regret theory and SSB theory is obscured by the fact that the former is formulated in terms of lotteries as acts, while the latter is formulated in terms of lotteries as probability distributions. Loomes and Sugden (1987) show that, for stochastically independent lotteries (where the set of states can be seen as a product of lottery-specific sets of states), generalized regret theory is equivalent to SSB theory. Again, the function $V^R$ becomes a particular example of the approach of Shafer (1974) for a space of lotteries.

Bordalo, Gennaioli, and Shleifer (2012, 2013) introduced *salience theory* by postulating a *symmetric* function $\sigma$, i.e. $\sigma(x,y) = \sigma(y,x)$ for all $x, y \in X \subseteq \mathbb{R}$, with the interpretation that for a lottery pair $(L^x, L^y)$, $\sigma(x_s, y_s)$ is the salience of the state $s$. This function is assumed to fulfill a number of properties capturing the idea of salience. In a "smooth" version of the theory, salience values are transformed through an increasing, real-valued function $f$ which preserves salience rankings as derived from $\sigma$, yielding[3]

$$q_s(L^x, L^y) = \frac{f(\sigma(x_s, y_s))}{\sum_{r \in S} f(\sigma(x_r, y_r))}.$$

The decision maker then attaches a value to lottery $L^x$ which depends on the alternative lottery $L^y$,

$$U^{ST}(L^x|L^y) = \sum_{s \in S} q_s(L^X, L^Y) u(x_s)$$

where $u$ is strictly increasing with $u(0) = 0$.

Although (smooth) salience theory appears functionally different from generalized regret theory and SSB models, it is worth observing that there is a relation. Under salience theory, a lottery $L^x$ is weakly preferred to a lottery $L^y$ if and only if $V^{ST}(L^x, L^y) \geq 0$, where

$$V^{ST}(L^x, L^y) = \sum_{s \in S} p(s) f(\sigma(x_s, y_s)) \left[u(x_s) - u(y_s)\right].$$

This already shows that regret theory is a further particular case of the approach of Shafer (1974) for a space of lotteries. Herweg and Müller (2021) further observe that the two-argument function on outcomes $w^{ST}$ defined by $w^{ST}(x,y) = f(\sigma(x,y))\left[u(x) - u(y)\right]$ is skew symmetric, and hence salience theory can be written in the same terms as generalized regret theory. Further, assuming continuity of $u$ and $f$, the assumptions of (smooth) salience theory imply those of generalized regret theory, that is, one can view salience theory as a particular case of the latter, and hence (for stochastically independent lotteries) as a particular case of SSB theory. Interestingly, the original regret theory of Loomes and Sugden (1982), which was a more specific model, turns out to be a particular case of salience theory if an additional, mild condition is imposed (Herweg and Müller, 2021, Theorem 2).

[3]Bordalo, Gennaioli, and Shleifer (2012) also provide a *rank-based* version of salience theory with similar insights. This version is analytically more tractable for specific applications, but creates discontinuities in valuations (Kontek, 2016).

All theories discussed above obviously allow for nontransitivities in lottery choice, since they are built upon the fundamental representation of Shafer (1974). That is, ultimately they provide a (structural, parametric) functional form for a function $V(\cdot,\cdot)$ defined on a specific space.[4]

## 3.2 Random Nontransitive Models

In this section, we extend the main result of Alós-Ferrer, Fehr, and Netzer (2021) to allow for nontransitivities. We consider abstract options, which could e.g. be themselves lotteries (this will be the case in our empirical analyses).

In an additive random utility model (McFadden, 1974, 2001), an agent is assumed to have an underlying utility function $u$ over a feasible set, but to be affected by random utility shocks. Thus, given a choice between two alternatives $x$ and $y$, realized utilities are $u(x) + \varepsilon_x$ and $u(y) + \varepsilon_y$, respectively, where $\varepsilon_x$, $\varepsilon_y$ are mean-zero random variables. Thus, a RUM generates choice probabilities, with the probability of $x$ being chosen when $y$ is also available given by

$$p(x,y) = \text{Prob}(u(x) + \varepsilon_x > u(y) + \varepsilon_y) = \text{Prob}(u(x) - u(y) + \varepsilon_x - \varepsilon_y > 0).$$

where tie-breaking conventions are irrelevant for continuously-distributed errors. Under specific assumptions on the distributions of the error terms, one obtains particular models, as the celebrated logit choice (Luce, 1959) or the classical probit choice (Thurstone, 1927). This general setting has become one of the dominant approaches in economics to model the fact that choice is empirically (and overwhelmingly) observed to be stochastic.[5]

Note that if the error term $\varepsilon_{xy} = \varepsilon_x - \varepsilon_y$ is assumed to be symmetrically distributed around zero, a preference for $x$ over $y$ is revealed if and only if $p(x,y) \geq 1/2$. A violation of transitivity in this framework thus consists of three (or more) alternatives $x, y, z$ such that $p(x,y) \geq 1/2$, $p(y,z) \geq 1/2$, and $p(z,x) > 1/2$. Hence

[4]It can be shown that generalized regret theory (and hence smooth salience theory) fulfill a stronger version of transitivity, called *dominance transitivity* by Diecidue and Somasundaram (2017): if $L^x$ strictly dominates $L^y$ (yields better outcomes for all states, and strictly better for at least some states) and the latter is preferred to $L^z$, then $L^x$ must be strictly preferred to $L^z$ (and analogously if $L^x$ is preferred to $L^y$ and the latter strictly dominates $L^z$). This rather strong condition seems to be the only systematic constraint on the kind of transitivity violations that these models can generate.

[5]The universe of random utility models comprehends also the class of random parameter models (e.g., Loomes and Sugden, 1998; Apesteguía and Ballester, 2018) as a special case where the distribution of errors is constrained by the structure of the family of utility functions, as well as drift-diffusion models (e.g., Ratcliff, 1978; Fudenberg, Strack, and Strzalecki, 2018; Baldassi et al., 2020). All these models assume exact functional forms mapping differences in utilities to error terms, which are valuable as structural assumptions but are in general not directly tested.

the literature tests for violations of Weak Stochastic Transitivity, which is defined as: if $p(x,y) \geq 1/2$ and $p(y,z) \geq 1/2$, then $p(x,z) \geq 1/2$.[6] However, since noise is not directly observable, the assumption of symmetric noise is untestable and might be unwarranted in general.

Alós-Ferrer, Fehr, and Netzer (2021) introduced a more general class of RUM models where error terms are modeled directly for utility differences, i.e. the realized utility difference given a choice $\{x,y\}$ is $u(x) - u(y) + \varepsilon_{x,y}$ for a mean-zero random variable $\varepsilon_{x,y}$ and hence

$$p(x,y) = \text{Prob}(u(x) - u(y) + \varepsilon_{x,y} > 0).$$

That work provided sufficient conditions on the distributions of response times conditional on each possible choice ($x$ or $y$ for a given pair $\{x,y\}$) which ensure that any RUM within a given class (defined by restrictions on the error terms, e.g. symmetry) which fits the data (in terms of choices and response times) reveals a preference for, say, $x$ over $y$, in the sense that $u(x) > u(y)$ for the underlying $u$. The importance of those results relies on the fact that they guarantee that an option is preferred to another for *any* utility function and *any* distribution of the error term that the analyst might consider, and hence the results are completely non-parametric and independent of functional forms. The message is that the properties of the empirical distribution of response times allow to recover the underlying preferences in random utility models without imposing any substantive assumptions on the distribution of random terms.

To allow for nontransitive preferences, we go one step forward and consider any skew-symmetric function $v : X^2 \mapsto \mathbb{R}$ (not necessarily arising from a utility function). That is, we consider models where noise is captured by mean-zero random variables $\varepsilon_{x,y}$ and choice probabilities are given by

$$p(x,y) = \text{Prob}(v(x,y) + \varepsilon_{x,y} > 0).$$

We remark at this point that our approach is agnostic with respect to whether decisions among lotteries are best represented by expected utility theory, prospect theory, or any other model generating preferences among lotteries. We merely test the class of models generating transitive choices, where the function above can be written as $v(x,y) = u(x) - u(y)$, against the class of models allowing for

[6] As discussed in Section 2, Regenwetter, Dana, and Davis-Stober (2010, 2011) and others have argued in favor of criteria other than WST to test for stochastic transitivity. However, WST remains a natural choice given our theoretical framework, and we will use it for ease of comparison to the literature.

nontransitivity lottery choices, where the function $v(x, y)$ cannot be written as a difference of utilities independently of the considered alternatives. The former class includes expected utility theory, rank-dependent utility theory, cumulative prospect theory, and others, while the latter includes generalized regret theory, salience theory, and SSB utility theory.[7]

To spell out the result, we need to define what we understand by a dataset. Given the set of alternatives $X$, denote by $C = \{(x, y) \mid x, y \in X, x \neq y\}$ the set of all binary choice problems, so $(x, y)$ and $(y, x)$ both represent the problem of choice between $x$ and $y$. Let $D \subseteq C$ be the set of choice problems on which we have data in the form of direct choices, assumed to be non-empty and symmetric, that is, $(x, y) \in D$ implies $(y, x) \in D$. A dataset (including response times) is modeled as follows (Alós-Ferrer, Fehr, and Netzer, 2021).

**Definition 1.** A *stochastic choice function with response times* (SCF-RT) is a pair of functions $(p, f)$ where

(i) $p$ assigns to each $(x, y) \in D$ a frequency $p(x, y) > 0$, with the property that $p(x, y) + p(y, x) = 1$, and

(ii) $f$ assigns to each $(x, y) \in D$ a strictly positive density function $f(x, y)$ on $\mathbb{R}_+$.

In an SCF-RT, $p(x, y)$ is interpreted as the proportion of the time that a decision maker chose $x$ when offered the binary choice between $x$ and $y$. The assumption that $p(x, y) > 0$ for all $(x, y) \in D$ implies that choice is noisy, that is, every alternative is chosen at least a small fraction of the time. The density $f(x, y)$ describes the distribution of response times conditional on the instances where $x$ was chosen in the binary choice between $x$ and $y$. The corresponding cumulative distribution function is denoted by $F(x, y)$. The following definition extends the concepts in Alós-Ferrer, Fehr, and Netzer (2021).

**Definition 2.** A *random nontransitive model with a chronometric function* (RNM-CF) is a triple $(v, \tilde{v}, r)$ where $v : X^2 \to \mathbb{R}$ is a skew-symmetric function and $\tilde{v} = (\tilde{v}(x, y))_{(x,y)\in C}$ is a collection of real-valued random variables, with each $\tilde{v}(x, y)$ having a density function $g(x, y)$ on $\mathbb{R}$, fulfilling the following properties:

(RNM.1) $\mathbb{E}[\tilde{v}(x, y)] = v(x, y)$,

[7] We remind the reader that our functions $u$ and $v$ are defined here on an abstract space. For a space of lotteries, $u$ might be expected utility and $v$ might be any of the functions $V^{SSB}$, $V^R$, $V^S$ described in Section 3.1.

(RNM.2) $\tilde{v}(x,y) = -\tilde{v}(y,x)$, and

(RNM.3) the support of $\tilde{v}(x,y)$ is connected.

Further, $r : \mathbb{R}_{++} \to \mathbb{R}_+$ is a continuous function that is strictly decreasing in $v$ whenever $r(v) > 0$, with $\lim_{v\to 0} r(v) = \infty$ and $\lim_{v\to\infty} r(v) = 0$.

A RUM-CF is a particular case of RNM-CF where the function $v$ is derived from a utility function, $v(x,y) = u(x) - u(y)$, and hence transitivity is guaranteed. The random variables $\tilde{v}(x,y)$ and their densities $g(x,y)$ capture noisy choice. Condition (RNM.1) requires that noise is unbiased (equivalent to assuming mean zero for an additive term $\varepsilon_{xy} = \tilde{v}(x,y) - v(x,y)$). Condition (RNM.2) reflects that the choice between $x$ and $y$ is the same as the choice between $y$ and $x$, and condition (RNM.3) is a regularity condition requiring connected support, i.e. without gaps. Last, $r$ represents the chronometric function, which maps realized values of $v$ into response times $r(|v|)$. Specifically, easier choices (where the value $\tilde{v}(x,y)$ is larger) are faster. This is in keeping with the interpretation that the function $v$ captures a strength of preference.

Given an RNM-CF $(v, \tilde{v}, r)$ and a pair $(x,y) \in C$, the random variable describing the response times predicted by the model conditional on $x$ being chosen over $y$ is given by

$$\tilde{t}(x,y) = r(|\tilde{v}(x,y)|),$$

conditional on $\tilde{v}(x,y) > 0$.

The results we seek will be in terms of preference revelation for *all* RNM-CFs which rationalize (explain) the data. The following definition pins down the formal meaning of the latter.

**Definition 3.** An RNM-CF $(v, \tilde{v}, r)$ *rationalizes* an SCF-RT $(p, f)$ if

(i) $p(x,y) = \text{Prob}[\tilde{v}(x,y) > 0]$ holds for all $(x,y) \in D$, and

(ii) $F(x,y)(t) = \text{Prob}[\tilde{t}(x,y) \leq t \mid \tilde{v}(x,y) > 0]$ holds for all $t > 0$ and all $(x,y) \in D$.

In other words, an RNM-CF (the model) rationalizes an SCF-RT (the data) if it reproduces both the choice frequencies and the conditional response time distributions in the latter. Obviously, fixing the set $D$, every RNM-CF generates an SCF-RT through the equations given in (i) and (ii) above, thus an alternative definition is that an RNM-CF rationalizes an SCF-RT if it coincides with the

SCF-RT generated by the former. We say that an SCF-RT is *rationalizable* if there exists an RNM-CF that rationalizes it. Note that an SCF-RT might be rationalizable by an RNM-CF even though it is not rationalizable by a RUM-CF.

The last definition captures preference revelation in a potentially nontransitive framework.

**Definition 4.** A rationalizable SCF-RT *reveals that $x$ is preferred to $y$* if all RNM-CFs that rationalize it satisfy $v(x, y) \geq 0$. It *reveals that $x$ is strictly preferred to $y$* if all RNM-CFs that rationalize it satisfy $v(x, y) > 0$.

The results in Alós-Ferrer, Fehr, and Netzer (2021) make use of the following technical concept. Given two cumulative distribution functions $G$ and $H$ on $\mathbb{R}_+$ and a constant $q \geq 1$, we say that $G$ *$q$-first-order stochastically dominates $H$* (also written $G$ $q$-FSD $H$) if

$$G(t) \leq q \cdot H(t) \text{ for all } t \geq 0.$$

If the inequality is strict for some $t$, then $G$ *strictly $q$-first-order stochastically dominates $H$* (written $G$ $q$-SFSD $H$). For $q = 1$, these concepts coincide with the standard notions of first-order stochastic dominance, but they are weaker when $q > 1$. Clearly, $q$-FSD implies $q'$-FSD whenever $q \leq q'$.

The following Theorem generalizes the main result of Alós-Ferrer, Fehr, and Netzer (2021) for the case of nontransitive preferences.

**Theorem 1.** *Consider random nontransitive models. A rationalizable SCF-RT $(p, f)$ reveals that $x$ is preferred to $y$ if $F(y, x)$ $q$-FSD $F(x, y)$, and that $x$ is strictly preferred to $y$ if $F(y, x)$ $q$-SFSD $F(x, y)$, for $q = p(x, y)/p(y, x)$.*

*Proof.* The proof is as the proof of Theorem 1 in Alós-Ferrer, Fehr, and Netzer (2021) replacing $u(x) - u(y)$ with $v(x, y)$. All arguments go through with the concepts amended as above. □

*Remark* 1. Note that the condition that $F(y, x)$ $q$-FSD $F(x, y)$ implies that $q \geq 1$ (e.g., by taking limits as $t \to \infty$) even if this were not stated as part of the definition. That is, if Theorem 1 reveals a (nontransitive) preference for $x$ over $y$, if follows that $p(x, y) \geq 1/2$, i.e. preferences cannot be revealed "against" choice frequencies, but choice frequencies do not imply preference revelation. This is important because most evidence for nontransitivities has been evaluated on the basis of Weak Stochastic Transitivity, which is stated in terms of choice frequencies.

Suppose that a dataset seems to point at nontransitive behavior, e.g. due to a violation of Weak Stochastic Transitivity. That is, the data identify a cycle of, say, three alternatives $x, y, z$ such that $p(x, y) \geq 1/2$, $p(y, z) \geq 1/2$, and $p(z, x) > 1/2$. While a researcher might take this as evidence of a transitivity violation, another researcher might argue that those population frequencies have arisen due to noise (as in a random utility model) even though underlying preferences are transitive. Until now, there was no way out of this debate, as there was no tool capable of determining whether an apparent violation of transitivity was due to noise or not.

Theorem 1 provides the missing tool. Suppose three alternatives $x, y, z$ build a violation of Weak Stochastic Transitivity for a given decision maker, as described above. If the dataset includes response times, one can apply the "Time Will Tell" (TWT) method derived from Theorem 1 to each of the pairs $(x, y)$, $(y, z)$, and $(x, z)$. In view of Remark 1, only two outcomes are possible. In the first case, preferences are revealed for all three pairs, which necessarily reveals a nontransitive cycle (except in the knife-edge case of full indifference). In this case, Theorem 1 shows that any model of choice explaining the observed data needs to entail a true nontransitive cycle, independently of the model of noise assumed (and, in particular, whether noise is symmetric or not). That is, in this case, a truly nontransitive cycle is revealed, which *cannot be due to noise.* In the second case preferences fail to be revealed for at least one of the pairs. In this case, the researcher is not entitled to conclude that the observed violation of Weak Stochastic Transitivity is actually due to a nontransitivity in underlying preferences; in other words, the observed violation might well be due to noise.

# 4 Empirical Evidence for Nontransitivity

## 4.1 Description of the Datasets

In this section we apply Theorem 1 to two existing datasets, both of which were specifically collected to study transitivity violations. The selected datasets, from Davis-Stober, Brown, and Cavagnaro (2015) (DSBC) and Kalenscher et al. (2010) (KTHDP), are ideal for our purposes because they include response times and every participant repeated every choice a reasonable number of times.

In the dataset of DSBC, $N = 60$ subjects made binary choices among different lotteries in a $2 \times 2$ within-subject design. Specifically, the experiment varied the display format of the lotteries (pies vs. bars) and whether participants faced a time constraint when making their choices or not (4 seconds vs. no time limit). The choice pairs were drawn from two sets of five lotteries each, with one lottery

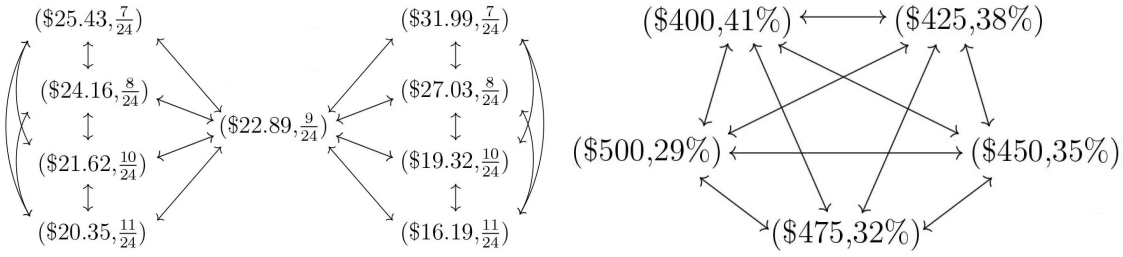


Figure 1: List of lotteries and implemented pairwise comparisons in Davis-Stober, Brown, and Cavagnaro (2015) (left) and Kalenscher et al. (2010) (right).

common to the two sets. All possible combinations of the lotteries within each set were implemented, giving rise to 20 distinct choice pairs (see Figure 1, left). Each of these pairs was repeated 12 times in each of the 4 possible conditions, for a total of $12 \times 4 \times 20 = 960$ choices per participant. Each participant took part in two sessions, with two (randomly allocated) combinations of time pressure and display format manipulations in each of them. Choices were incentivized (one decision from each condition was randomly selected and paid, in addition to a show-up fee).

In the dataset of KTHDP, $N = 30$ subjects made binary choices among five different lotteries.[8] All combinations of the lotteries where implemented (see Figure 1, right). Each of the 10 resulting choice pairs was repeated 20 times, for a total of 200 trials per participant. Participants needed to decide within 4 seconds, with missed time limits resulting in a missed trial. Each participant took part in a single, individual-level session while being scanned in an fMRI machine. Choices were incentivized (with dummy dollars translated into Euro with a conversion rate of 100:1), with one randomly-selected decision paid in addition to a show-up fee.

In addition to the presence of repetitions, the measurement of response times, and the fact that they were collected to study transitivity violations, the two datasets are also interesting for other reasons. First, all lotteries involve only one non-zero outcome and hence can be presented with only two variables (a single outcome and its probability). This makes alternatives easy to compare for participants. Second, all magnitudes in each of the experiments are comparable (without extreme differences), hence mitigating possible concerns regarding range or outlier effects. Third, none of the lotteries involves probabilities close to zero or one, which are known to bias behavior.

[8]Further 240 filler lotteries were used, but they all were paired in a way which involved dominated choices, and hence are not interesting for our purposes.

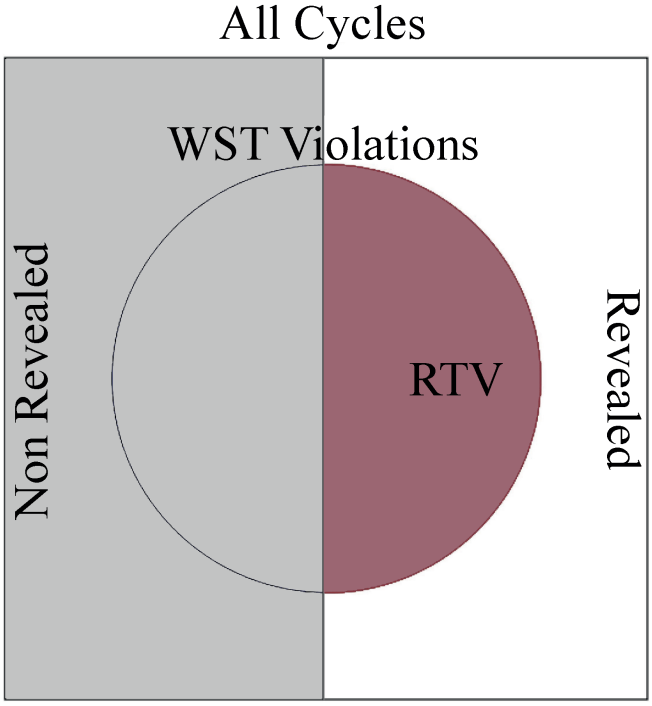


Figure 2: If the subset of WST violations (subset of all cycles in the circle) are due to noise, then the set of revealed violations (subset in red) should be empty. However, if transitivity violations are not noise, then this subset should be non-empty. In practice, this subset must coincide with set of all WST violations where preferences are revealed for each involved pair.

## 4.2 Strategy of Analysis and Preliminaries

We now investigate transitivity violations in the two datasets. A *Revealed Transitivity Violation* (RTV) exists in the data whenever application of Theorem 1 reveals a cycle with $x_1 \succeq x_2 \succeq \ldots \succeq x_n$ and $x_n \succ x_1$. An RTV reveals a nontransitivity which cannot be explained by noise, and in this sense disentangles noise from true transitivity violations. Since, within the universe of RNMs, preferences revealed by our method are independent of the form of noise assumed, we conclude that transitivity violations which involve only revealed preferences cannot be due to any form of noise or because of the specific function $v$ that one assumes.

Up to now, the empirical literature has predominantly looked at violations of *Weak Stochastic Transitivity* (WST) to study transitivity violations. This property states that for all $x_1, x_2, x_3$ such that $p(x_1, x_2) \geq 1/2$ and $p(x_2, x_3) \geq 1/2$, it must follow that $p(x_1, x_3) \geq 1/2$. Other concepts of transitivity in a stochastic setting exist, as e.g. strong stochastic transitivity (where the implication is that $p(x_1, x_3) \geq \max\{p(x_1, x_2), p(x_2, x_3)\}$) or moderate stochastic transitivity (which replaces the

maximum with the minimum in the previous implication). See Fishburn (1998) for an overview.

The concept of RTV is more stringent than violations of WST. If a nontransitive cycle $x_1 \succeq x_2 \succeq x_3 \succ x_1$ is revealed by an application of Theorem 1, it follows from Remark 1 that this cycle also entails a WST violation. Hence, the concepts are naturally nested, that is, every RTV is necessarily a WST violation. The question we ask is what is the proportion of empirical WST violations where the researcher is actually justified to conclude that a transitivity violation actually exists and is not simply due to behavioral noise.

The intuition behind the relation between WST violations and our analysis is sketched in Figure 2. Fix a cycle of alternatives, $(x_1, x_2, \ldots, x_n, x_{n+1} = x_1)$. We apply Theorem 1 to the data for every binary choice along the cycle, $\{x_i, x_{i+1}\}$, $i = 1, \ldots, n$. If any of the preferences along this cycle is not revealed (neither $x_i \succeq x_{i+1}$ nor $x_{i+1} \succeq x_i$, as in the left-hand-side part of the figure), then no conclusion can be drawn as to whether the cycle entails a transitivity violation or not. However, data can still show a WST violation. In that case, the researcher is not entitled to conclude that a true transitivity violation exists, as the choice proportions might be due to noise. If all preferences along the cycle are revealed after application of Theorem 1, those might build a transitivity violation (an RTV) or not. We know from Remark 1 that all RTVs must be violations of WST. Conversely, if all preferences along a cycle violating WST are (strictly) revealed, again by Remark 1 the cycle must in practice be an RTV. For, if a preference between $x$ and $y$ is revealed and $p(x, y) > 1/2$, only a preference of $x$ over $y$ can be revealed.[9]

If choices were always transitive, empirically-observed WST violations would be due to noise. Then, once we identify which preferences are revealed, the subset of RTVs should be empty. On the other hand, if transitivity violations are not due to noise, then the subset of cycles where all preferences are revealed should still contain violations of transitivity, i.e. RTVs. The size of this set relative to the size of the set of cycles involving WST violations (and where all preferences along the cycle are revealed) quantifies how accurate WST actually is in detecting true transitivity violations, conditional on preferences being revealed. The size of the

[9]In principle, it is possible that $p(x, y) > 1/2$ but the TWT method only reveals a weak preference, which would make it possible to have a WST violation which cannot be concluded to be an RTV (instead of an indifference cycle). It is also possible that a WST violation involves $p(x, y) = 1/2$ and the TWT method reveals a strict preference either way, hence allowing for WST violations where all preferences are (even strictly) revealed but a nontransitive cycle does not arise. In practice, such knife-edge cases are empirically rare and they never occurred in our data.

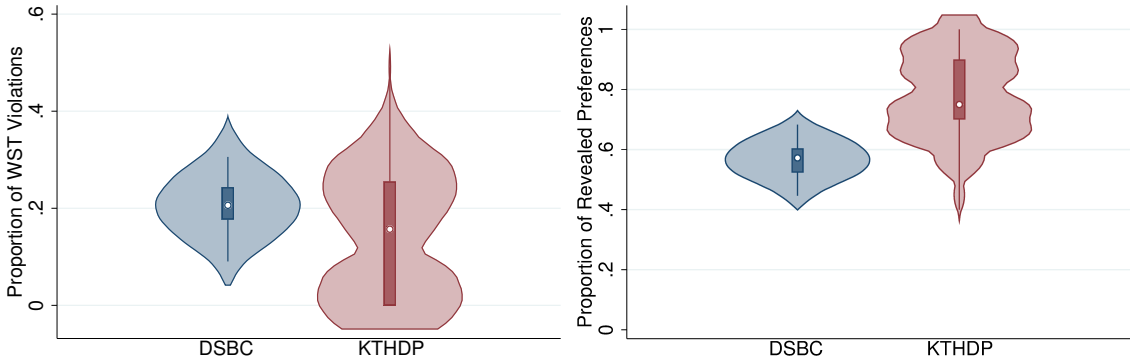


Figure 3: Distribution of the individual proportions of WST (on the left) and of revealed preferences (on the right). Violin plots show the median, the interquartile range and the 95% confidence intervals as well as rotated kernel density plots on each side.

set of RTVs relative to the size of all WST violations quantifies the accuracy of taking the latter as a proxy of true transitivity violations.

As explained above, we aim to compare revealed nontransitivities according to Theorem 1 with violations of WST. For this analysis to be feasible, two preconditions must be fulfilled. First, enough WST violations should be present in the data. Second, a relatively large proportion of the observed choices should lead to the underlying preferences being revealed by Theorem 1. Figure 3 illustrates that both preconditions are met. The left-hand side of this figure displays violin plots for the subject-level proportion of WST violations (the differences across datasets are presumably due to differences in stimuli and experimental implementation). For DSBC, we observe an average of 20.77% WST violations (median 20.61%, SD=5.28, min 9.04%, max 34.57%), while in KTHDP the average is 15.42% (median 15.69%, SD=13.93, min 0.00%, max 49.02%). These proportions are roughly representative of results in the literature, and indicate a sizeable percentage of transitivity violations if WST is used as a criterion.

The right-hand-side of Figure 3 illustrates how often application of Theorem 1 reveals preferences. That is, for every potential cycle and every binary choice along the cycle, we compute the choice proportions and the response time densities and check whether the condition in Theorem 1 holds.[10] For DSBC, the average percentage of choices at the subject level for which the method reveals preferences

[10]To reveal preferences using the TWT method, we need to estimate the density of the distribution of response times. As in Alós-Ferrer, Fehr, and Netzer (2021), the kernel density estimates were performed in *Stata* using the *akdensity* function, which delivers CDFs as output. We estimate the distribution of log-transformed response times to avoid boundary problems. The estimates use an Epanechnikov kernel with optimally chosen non-adaptive bandwidth. For the case where some choice is made only one out of the total number of repetitions (only a single

is 56.67% (median 57.22%, SD= 6.15, min 44.66%, max 68.31%), while for KTHDP is 77.00% (median 75.00%, SD=15.57, min 40.00%, max 100.00%). Thus, in our datasets, the method reveals preferences often enough for an analysis of revealed nontransitivities to be conducted.

*Remark* 2. For DSBC participants, we find no differences in the proportion of revealed preferences depending on whether subjects were under time pressure or not (56.13% vs. 57.16%; WRS, $N = 60$, $z = -0.942$, $p = 0.3505$). This is important, as it suggests that even though the method relies on response times, its capacity to reveal preferences is not affected by (reasonable) time limits, and hence it is robust with respect to such manipulations.

## 4.3 Revealed Transitivity Violations

We now turn to our main analysis. Say that a cycle of alternatives

$$(x_1, x_2, \ldots, x_n, x_{n+1} = x_1)$$

is a *revealed cycle* if all preferences along the cycle are revealed, i.e. the method reveals either $x_i \succeq x_{i+1}$ or $x_{i+1} \succeq x_i$, for all $i = 1, \ldots, n$. The proportion of revealed cycles is obviously smaller than the proportion of choices for which preferences are revealed, since all preferences along a cycle must be revealed for the cycle to be revealed. For DSBC, 20.82% of cycles are revealed (median 22.08%, SD=7.61, min 0.00%, max 32.08%), and the number is 54.25% (median 60.00%, SD=25.00, min 0.00%, max 100.00%) for KTHDP.

Figure 4 illustrates how the stylized partition of data sketched in Figure 2 looks like for the two actual datasets. We compute the proportion of all WST violations where the cycle is revealed and check that they are indeed RTVs. Recall that every RTV is a WST violation, and, except for knife-edge cases, every WST violation where preferences are revealed is an RTV. Indeed, in both datasets, every single WST violation for a revealed cycle is also an RTV.

We obtain that, on average across subjects, 19.24% of all WST violations are actually RTVs for DSBC (median 17.71%, SD=9.56, min 4.35%, max 43.42%). The average is 39.58% for KTHDP (median 29.41%, SD=32.60, min 0.00%, max 100.00%). That is, when using WST to evaluate whether people violated transitivity, in about a third of cases the data provide enough evidence to support the

response time is available) an optimal bandwidth cannot be determined endogenously, so we set it to 0.1, yielding a distribution function close to a step function at the observed response time.

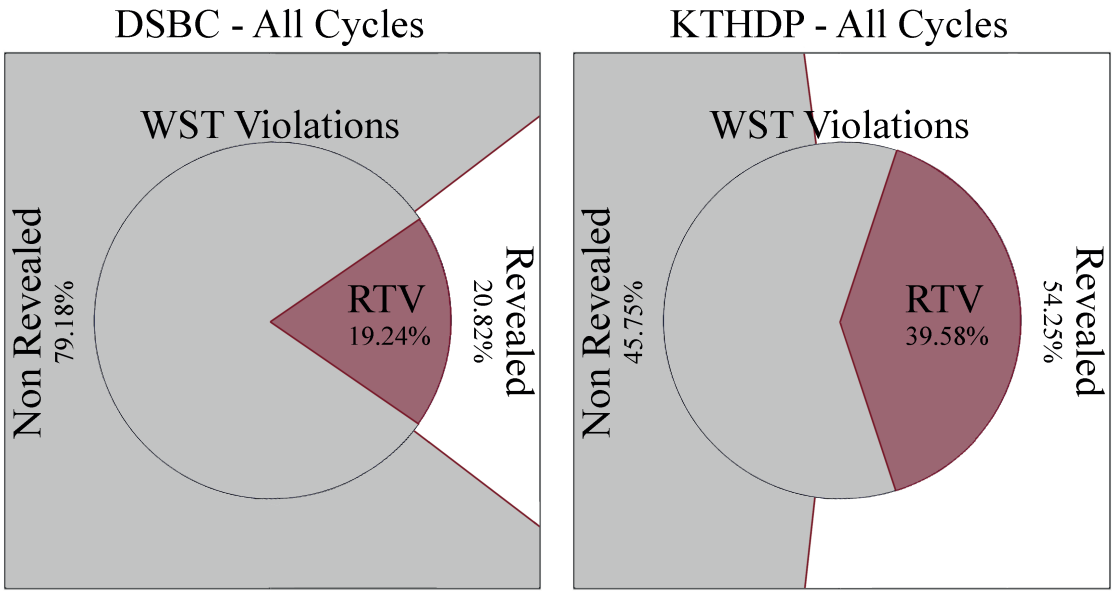


Figure 4: Proportion of all WST violations where the cycle is revealed or not for DSBC (on the left) and KTHDP (on the right). All WST violations where the cycle is revealed are RTVs.

statement that these choices are not just behavioral noise. In the rest of the cases, the researcher is not entitled to state that WST violations are not just noise.

In other words, for 19.24% of all WST violations for DSBC, and 39.58% for KTHDP, application of Theorem 1 reveals a transitivity violation which cannot be due to noise. In particular, we can conclude that true transitivity violations indeed exist within the universe of RNM models independently of any assumption made on the distribution of noise or the underlying value function. At the same time, for most of the observed WST violations, it cannot be discarded that they are simply due to some sort of underlying noise. That is, in the majority of cases which the literature has identified as WST violations, data might not actually allow to conclude that these reflect actual transitivity violations.

As a percentage of all decisions, the set of transitivity violations (RTVs) is comparatively small. The individual proportion of revealed transitivity violations over all cycles in DSBC is 4.03% (median 3.72%, SD=2.27, min 0.53%, max 10.32%) and 6.47% in KTHDP (median 1.96%, SD=8.40, min 0.00%, max 27.45%).

The message which arises from our analysis is two-fold. First and foremost, our approach identifies transitivity violations which *cannot* be explained by noise (at least within the framework of RNMs), and hence the set of violations we identify stand on conceptually solid ground as a demonstration that nontransitivities in the data do occur. Second, however, we conclude that evidence for transitivity violations, as a percentage of all decision cycles, is smaller than one would conclude

by using previous measures, in particular violations of WST, in the sense that, for many of those, it is unwarranted to conclude that a true transitivity violation has been discovered. We now turn to a more detailed comparison.

## 4.4 Comparison Between RTVs and WST Violations

We would like to quantify the size of the set of transitivity violations at the individual level, and compare it to previous measurements using WST. Since the number of RTVs for a given subject is necessarily smaller than the individual number of WST violations (Remark 1), we quantify the proportion of RTVs in relation to cycles with revealed preferences only. That is, we compute the subject-level number of RTVs divided by the total number of cycles for which preferences are revealed along the entire cycle. This proportion is not mechanically related to the proportion of WST violations, and hence this procedure allows a fair comparison of the magnitudes of transitivity violations as suggested by RTV and WST.

Figure 5 plots the distribution of subject-level proportions of RTV over all *revealed* cycles, that is, excluding cycles where preferences were not revealed, for both datasets. In particular, if transitivity violations would mainly arise from choices which are not revealed, we should see a sharp decrease in the proportion of transitivity violations according to RTV when computed in this way (since non-revealed cycles are excluded), when compared to WST violations. On the contrary, if transitivity violations are orthogonal to whether preferences are revealed by Theorem 1 or not, the overall proportion of transitivity violations according to WST and to RTV computed in this way should be unaffected.

The individual proportion of revealed transitivity violations in DSBC is 19.24%, compared to a proportion of 20.77% of WST violations for the overall sample. The difference is small, and a Wilcoxon Rank-Sum test reveals no significant differences at the 5% level ($N = 60$, $z = -1.811$, $p = 0.0705$). In KTHDP the proportion of RTVs is 13.83%, compared to a 15.42% of WST violations for the overall sample. However, again there are no significant differences at the 5% level (WRS, $N = 29$, $z = -1.847$, $p = 0.0657$). Hence, the evidence is aligned with the interpretation that transitivity violations might be orthogonal to whether preferences are revealed by Theorem 1 or not. However, of course, this is just suggestive evidence and one cannot conclude that WST violations where preferences are not revealed are actually transitivity violations.

*Remark* 3. In Appendix B.1 we take advantage of the manipulations in DSBC to further investigate the robustness of the results. In that experiment, participants faced lotteries in two different graphical formats, both with and without

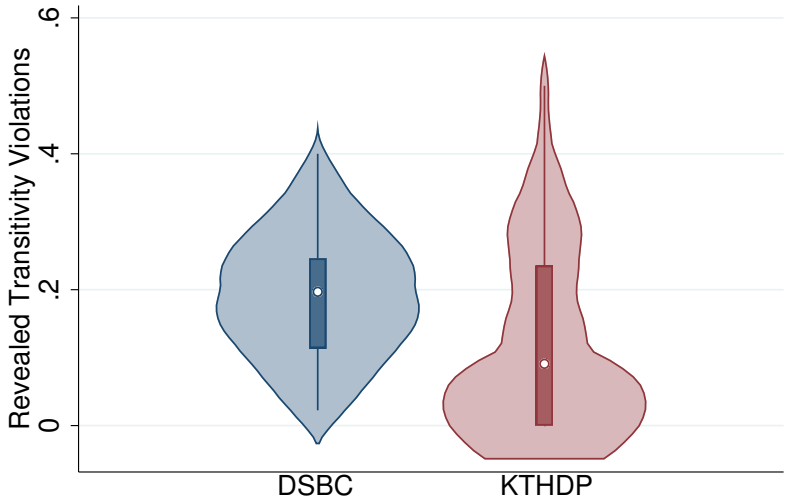


Figure 5: Distribution of the individual proportions of RTV over all cycles where all preferences are revealed. Violin plots show the median, the interquartile range and the 95% confidence intervals as well as rotated kernel density plots on each side.

time pressure. The results reported above are qualitatively unchanged by either manipulation.

*Remark* 4. Figures 3 and 5 show clear differences between the two datasets. Participants in DSBC display limited heterogeneity, with unimodal, relatively concentrated distributions of WST violations and RTVs. In contrast, in KTHDP there seem to be more heterogeneity across participants, with a more disperse distribution. In KTHDP, the authors report that participants were unaware of having made intransitive choices, but some reported following heuristic rules of behavior which would indeed produce transitivity violations within the context of the experiment. For our purposes, the fact that we obtain comparable results from two radically different samples and experiments strengthens our conclusions.

## 4.5 Characteristics of Nontransitive Cycles

Our analysis above shows the existence of transitivity violations which are not due to noise. A natural question is whether specific collections of lottery choices give rise to such violations often. To answer this question, we reanalyze the data taking individual cycles as the unit of observation. That is, in each dataset and for each cycle of alternatives, we compute the percentage of participants who display either a WST violation or an RTV. The left-hand panel of Figure 6 shows the distribution of the proportion of participants displaying WST violations across

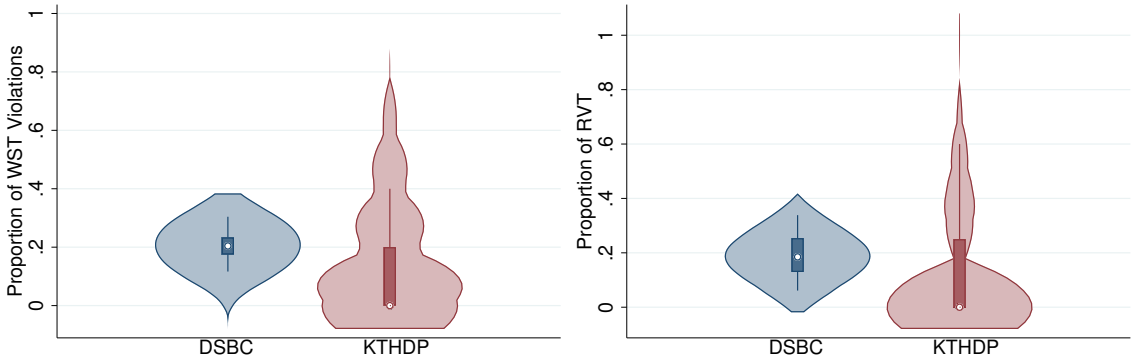


Figure 6: Distribution of the proportion of subjects displaying WST violations (on the left) and RTVs (on the right) per each cycle.

cycles (DSBC: mean 20.56%, median 21.67%, SD=6.51, min 0.00%, max 36.67%; KTHDP: mean 15.42%, median 0.00%, SD=20.03, min 0.00%, max 80.00%), while the right-hand panel represents the distribution of the proportion of participants displaying RTVs across cycles, computed over all participants for which the cycle was revealed (DSBC: mean 18.84%, median 17.16%, SD=11.50, min 0.00%, max 58.33%; KTHDP: mean 11.22%, median 0.00%, SD=20.55, min 0.00%, max 100.00%).[11] The data hence reveals heterogeneity across cycles, that is, some cycles involve nontransitive choices for a sizeable part of the experiment's participants, while others involve next to no violations.

To single out which constellations of choices produce a particularly large proportion of violations, we then look at the cycles which entail the most transitivity violations. Table 1 lists the ten cycles (for both datasets) with the largest proportion of RTVs, computed as the percentage of people for which the cycle was revealed who displayed an RTV. For DSBC, those range from 48% to 58%, and all of them correspond to WST violations for at least a quarter of the sample. Notably, all ten cycles involve just the five following lotteries (out of the nine in the experiment), which correspond to the left-hand subset in Figure 1(left).

$$x_1 = \left(\$25.43, \frac{7}{24}\right),\ x_2 = \left(\$24.16, \frac{8}{24}\right),\ x_* = \left(\$22.89, \frac{9}{24}\right),$$
$$x_3 = \left(\$21.62, \frac{10}{24}\right), x_4 = \left(\$20.35, \frac{11}{24}\right)$$

The fact that the most common transitivity violations in DSBC all involve the left-hand subset in Figure 1(left), and none of them involves the lotteries in the

[11] Note that for DSBC the average is computed over $N = 60 \times 4$ observations, as each participant made the same choices in four different conditions.

Table 1: The ten cycles in DSBC and KTHDP with the most transitivity violations. The second column indicates the proportion of experimental participants displaying an RTV for the cycle in the first column, computed over all participants for which the cycle was revealed (numbers in brackets indicate how the proportion is computed). The third column indicates the proportion of participants (out of $4 \times 60$ for DSBC, out of 30 for KTHDP) displaying a WST violation for the cycle.

| Cycle | People with RTV | People with WST |
|---|---|---|
| DSBC | | |
| $x_* \succ x_4 \succ x_3 \succ x_*$ | 58.33% (28/48) | 30.00% (72) |
| $x_* \succ x_4 \succ x_3 \succ x_2 \succ x_*$ | 54.55% (24/44) | 35.00% (84) |
| $x_* \succ x_1 \succ x_4 \succ x_3 \succ x_*$ | 50.00% (12/24) | 25.00% (60) |
| $x_* \succ x_4 \succ x_2 \succ x_3 \succ x_*$ | 53.85% (28/52) | 31.67% (76) |
| $x_* \succ x_2 \succ x_3 \succ x_4 \succ x_*$ | 57.14% (32/56) | 33.33% (80) |
| $x_* \succ x_2 \succ x_3 \succ x_1 \succ x_4 \succ x_*$ | 47.62% (40/84) | 36.67% (88) |
| $x_* \succ x_4 \succ x_1 \succ x_3 \succ x_2 \succ x_*$ | 50.00% (24/48) | 35.00% (84) |
| $x_* \succ x_1 \succ x_4 \succ x_2 \succ x_3 \succ x_*$ | 50.00% (12/24) | 26.67% (64) |
| $x_* \succ x_4 \succ x_1 \succ x_2 \succ x_3 \succ x_*$ | 53.85% (28/52) | 35.00% (84) |
| $x_* \succ x_4 \succ x_2 \succ x_1 \succ x_3 \succ x_*$ | 57.14% (32/56) | 33.33% (80) |
| KTHDP | | |
| $y_2 \succ y_4 \succ y_5 \succ y_2$ | 66.67% (12/18) | 40.00% (12) |
| $y_2 \succ y_3 \succ y_5 \succ y_4 \succ y_2$ | 100.00% (6/6) | 60.00% (18) |
| $y_3 \succ y_1 \succ y_2 \succ y_4 \succ y_3$ | 66.67% (12/18) | 60.00% (18) |
| $y_3 \succ y_4 \succ y_5 \succ y_1 \succ y_3$ | 66.67% (12/18) | 40.00% (12) |
| $y_4 \succ y_1 \succ y_2 \succ y_3 \succ y_4$ | 66.67% (12/18) | 60.00% (18) |
| $y_1 \succ y_4 \succ y_3 \succ y_2 \succ y_5 \succ y_1$ | 75.00% (9/12) | 30.00% (9) |
| $y_3 \succ y_1 \succ y_2 \succ y_4 \succ y_5 \succ y_3$ | 66.67% (12/18) | 60.00% (18) |
| $y_3 \succ y_4 \succ y_5 \succ y_1 \succ y_2 \succ y_3$ | 66.67% (12/18) | 40.00% (12) |
| $y_4 \succ y_1 \succ y_2 \succ y_3 \succ y_5 \succ y_4$ | 66.67% (12/18) | 60.00% (18) |
| $y_4 \succ y_5 \succ y_2 \succ y_1 \succ y_3 \succ y_4$ | 66.67% (12/18) | 40.00% (12) |

right-hand set, is particularly revealing. The differences in outcomes across similar lotteries in the left-hand set are noticeably larger (between \$3.13 and \$4.96) than those for the other set (all \$ 1.27), while differences in probabilities are always 1/24 in both sets. That is, the most frequent nontransitivities involve choices whose evaluations are presumably closer, i.e. such that the strength of preference is smaller. If one used WST or a similar measure as a criterion for detecting nontransitivites, standard psychometric effects (error rates are larger for closer valuations) would suggest that the increase in nontransitivities is merely due to increased noise. However, our approach through RTVs has disentangled preferences from noise. Thus, the data suggests that the increase in nontransitivities is due to the fact that evaluations are close, but not because this results in noisier choices. Rather, it appears that empirical transitivity violations are more frequent

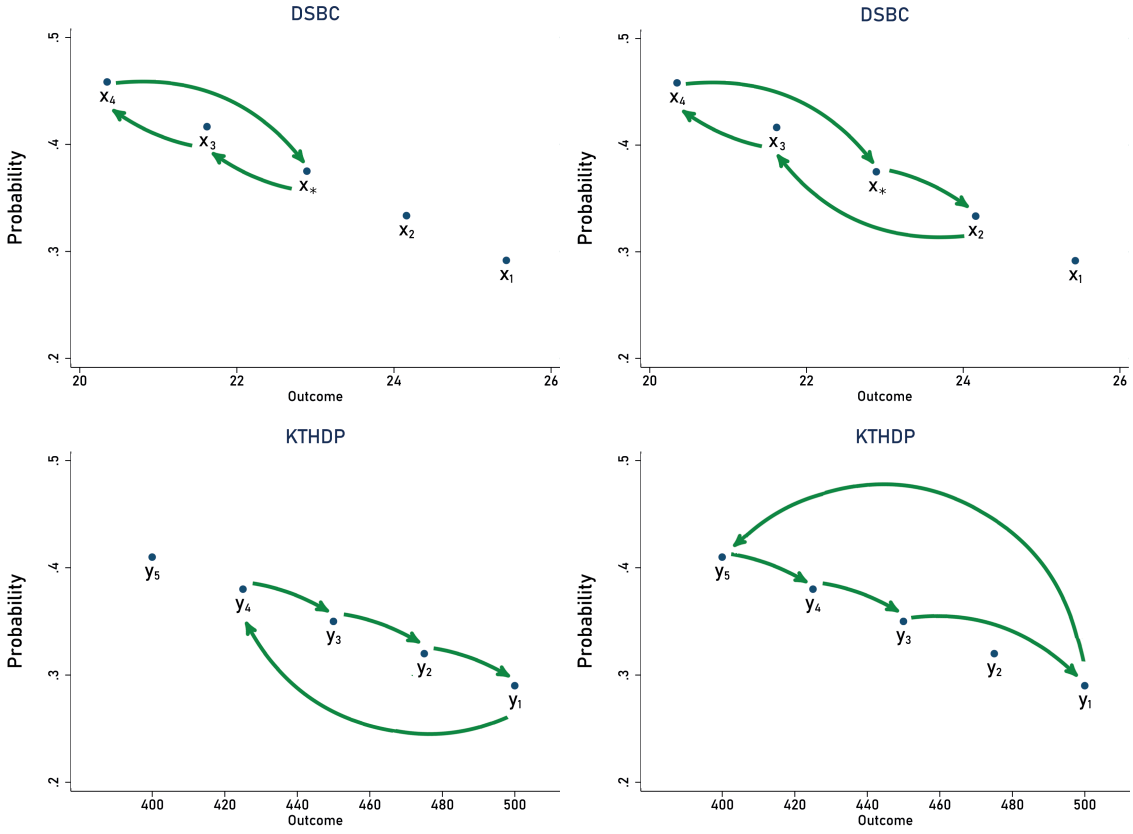


Figure 7: Graphical representation of some of the most common cycles in the datasets. All lotteries have a single non-zero outcome, depicted in the (outcome, probability) space. Arrows indicate preference, i.e. $x \to y$ means $y \succ x$. The two upper pictures are from the DSBC data, the two lower ones from KTHDP.

when they result from a gradual chain of small changes in the options. Specifically, many of the examples in Table 1 suggest that small tradeoffs, which are possible when lottery attributes are close enough, do not scale up monotonically. For example, consider the shortest cycle for DSBC in Table 1, which is also the one with the largest proportion of RTV violations, $x_* \succ x_4 \succ x_3 \succ x_*$. Twice along this cycle ($x_4 \succ x_3 \succ x_*$), the decision maker accepts a one-step decrease in monetary payoff (\$1.27) in exchange for a one-step increase in probability (1/24). Then, however, the same decision maker accepts a two-steps decrease in probability (2/24) in exchange for a two-step increase in monetary payoff (\$2.54). The exact same phenomenon appears in the cycles $x_* \succ x_4 \succ x_3 \succ x_2 \succ x_*$, $x_* \succ x_2 \succ x_3 \succ x_4 \succ x_*$, and (rewritten) $x_1 \succ x_4 \succ x_3 \succ x_* \succ x_1$, with three one-step tradeoffs being reversed by a three-step tradeoff in the opposite direction, and similar but more complex patterns can be seen in the longer cycles. The two top panels of Figure 7 give a graphical representation of two of these examples.

For KTHDP, the proportion of RTVs among revealed cycles for the ten topmost ones is always above two thirds. corresponding to between 40% and 60% WST violations in the overall sample. The cycles involve all five lotteries in KTHDP,

$$y_1 = (\$500, 0.29)\,,\ y_2 = (\$475, 0.32)\,,\ y_3 = (\$450, 0.35)\,,$$
$$y_4 = (\$425, 0.38)\,,\ y_5 = (\$400, 0.41)$$

The same phenomenon is observed in several of the KTHDP cycles. For example, in the cycle $y_4 \succ y_1 \succ y_2 \succ y_3 \succ y_4$, three times in a row the decision maker accepts a one-step reduction in probability (0.03) in exchange for a one-step increase in monetary payoff ($25), but then undoes it by accepting a three-step reduction in monetary payoff ($75) in exchange for a three-step increase in probability (0.09). A similar pattern can be seen in the cycle $y_3 \succ y_4 \succ y_5 \succ y_1 \succ y_3$, and similar phenomena appear in several of the longer cycles. The two bottom panels of Figure 7 give a graphical representation of two of these examples.

# 5 Predicting Choices Out of Sample in Spite of Nontransitivities

Our results raise a natural question. So far, we have provided conclusive evidence that a percentage of decisions under risk (in the datasets we analyze) entail transitivity violations which cannot be ascribed to noise. On the other hand, the percentage of such transitivity violations is naturally smaller than previously assumed in the literature if one takes WST violations as the criterion, as those include observations that might be just due to noise. The natural follow-up question is whether transitivity violations fundamentally impair our capacity to forecast or predict decisions out of sample, or rather can just be ignored for these purposes, as yet another factor making predictions imperfect. Strictly speaking, our results reject the hypothesis of transitivity, and standard prediction methods (based, e.g., on the estimation of underlying utilities) do assume transitivity. However, if the actual number of effective violations is small, this (fundamental) theoretical problem might not pose unsurmountable empirical difficulties.

To address this question, we used the datasets of DSBC and KTHDP to perform out-of-sample prediction exercises. In doing so, we also compared the performance of parametric and non-parametric prediction methods. Parametric methods typically entail the estimation of a utility function (with a pre-specified shape, say CARA or CRRA) within the context of a random utility model (Anderson, Thisse,

and De Palma, 1992; McFadden, 2001) or a random parameter model (Loomes and Sugden, 1998; Apesteguía and Ballester, 2018), with additional, specific functional assumptions on the shape of the noise (e.g., logit or probit models). Appendix A briefly summarizes the (standard) microeconometric approach we followed.

In contrast, we also considered prediction methods derived from additional results in the TWT approach (Alós-Ferrer, Fehr, and Netzer, 2021). Those entail the non-parametric estimation of a (transitive) preference using response times and choice frequencies, which is possible under the additional (also non-parametric) assumption that the noise term is symmetric.[12] This of course entails two possibly-unwarranted assumptions, the transitivity of the underlying preference and the symmetry of the noise. Under symmetric noise, in this case, $p(x,y) > p(y,x)$ reveals a strict preference for $x$ over $y$. To obtain out-of-sample predictions, the idea is to triangulate a preference indirectly through comparisons with a reference option. The rough intuition is that, if an option $a$ is preferred to $x^*$ with fast response times, this preference is relatively strong. If another option $b$ is preferred to the same $x^*$ with slow response times, this preference is relatively weak. Even though no conclusion follows from transitivity (as both $a$ and $b$ are preferred to $x^*$), the cardinality embodied in response times should allow to conclude that $a$ is preferred to $b$. Theorem 2 in Alós-Ferrer, Fehr, and Netzer (2021) shows that, however, this intuition is elusive, and the meaning of "fast" and "slow" is subtle. Specifically, for each $(x,y) \in D$ with $p(x,y) > p(y,x)$, define $\theta(x,y)$ as the $1/2p(x,y)$-percentile of the response time distribution of $x$, i.e., $F(x,y)(\theta(x,y)) = \frac{1}{2p(x,y)}$ The quantity $\theta(x,y) > 0$ combines information about choice probabilities and response times, that is, it corresponds to a different percentile for each choice pair. Once one replaces "fast or slow response time" with $\theta(x,y)$, the result fully captures the intuition above. We restate it here spelling out all implicit conditions in Alós-Ferrer, Fehr, and Netzer (2021) for convenience.

**Theorem 2** (Theorem 2, Alós-Ferrer, Fehr, and Netzer, 2021)**.** *Within the class of symmetric RUM-CFs, a rationalizable SCF-RT reveals a preference for $x$ over $y$, where $(x,y) \in C \setminus D$, if there exists $x_* \in X$ such that $(x,x_*),(y,x_*) \in D$ and*

- *either $p(x,x_*), p(y,x_*) > \frac{1}{2}$ and $\theta(x,x_*) \leq \theta(y,x_*)$*
- *or $p(x_*,x), p(x_*,y) > \frac{1}{2}$ and $\theta(x_*,x) \geq \theta(x_*,y)$*

*and it reveals a strict preference if the inequalities are strict.*

[12] Noise in a RNM-CF is symmetric if for each $(x,y) \in C$ and all $\delta \geq 0$, $g(x,y)(v(x,y)+\delta) = g(x,y)(v(x,y)-\delta)$.

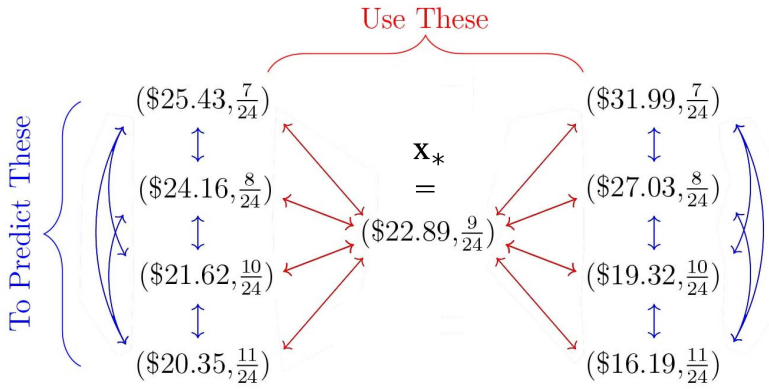


Figure 8: Assuming symmetric noise in a dataset with a reference option ($x_1$) we can use all comparisons with this reference (in red) to predict out-of-sample all other choices (in blue).

Hence, by fixing a reference option $x_*$, one can derive a full (transitive) preference among all alternatives which have been compared to $x_*$ in a dataset including response times. The price to pay is, as commented above, twofold. First, one assumes symmetric noise. Second, transitivity of the underlying preference relation is taken for granted. Of course, we know that both of these assumptions are violated (see Alós-Ferrer, Fehr, and Netzer, 2021 for a discussion of the assumption of symmetric noise), i.e. the model underlying this prediction is incorrect. The question, however, is not whether the model (transitive preferences and symmetric noise) is entirely correct, but rather how useful it is for predictive purposes, or, in other words, how severe are the consequences of transitivity violations (and noise asymmetry) for prediction.

As Figure 8 shows, the dataset of DSBC has a particular structure which is especially interesting for our purposes. All lotteries were repeatedly compared to a specific one (denoted $x_*$ in the figure, where these comparisons are highlighted in red). Theorem 2 above then allows to estimate a full preference relation using just those decisions. In particular, the estimated preferences allow to predict the choices among other lotteries. The dataset also includes choices within the left-hand- and right-hand subsets (highlighted in blue in Figure 8), and hence we can test the predictions. That is, we can use the choice frequencies and response times of the first (red) type of choices in order to predict all other comparisons (in blue) out-of-sample.

As a comparative benchmark, we use a standard econometric approach mimicking the procedure described above. We estimated risk attitudes, separately for each individual, based only on the choices which involved $x_*$ (see Appendix A for details on the estimation procedure). This is the same subset we used for the

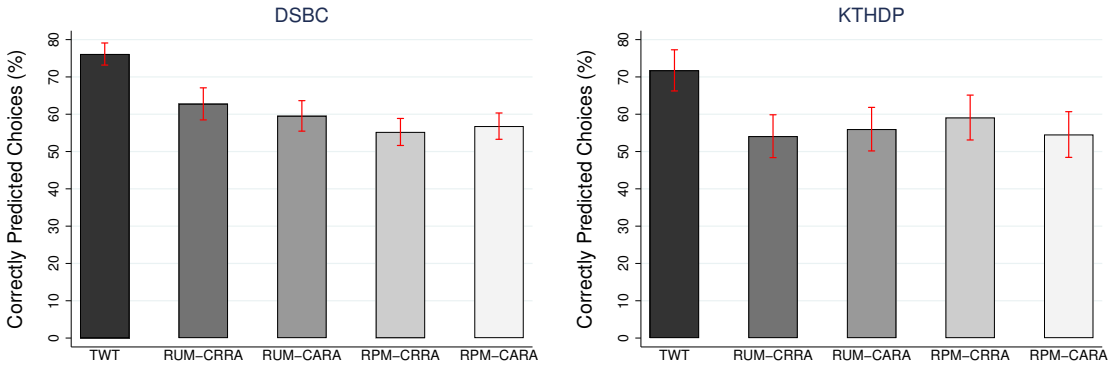


Figure 9: Proportion of out-of-sample correctly predicted choices for RUM/RPM and TWT for Davis-Stober, Brown, and Cavagnaro (2015) (on the left) and Kalenscher et al. (2010) (on the right) across different utility functions (CRRA vs. CARA). 95% confidence intervals are represented in red.

TWT exercise, in red in Figure 8. We then used this individual estimate to predict all other choices not involving the lottery $x_*$ (in blue; again, this is the same set as in the TWT application). In order to do this, however, we need structural assumptions on both the utility function and the noise term. As frequently done in the literature, we assume a random utility model with a CRRA utility function, and conduct a robustness check by repeating the analysis with a CARA function. We further repeat the analysis assuming a random parameter model (with either CRRA or CARA functions) instead.[13]

The dataset of KTHDP does not have the structure of DSBC. In this dataset, all binary choices among five different lotteries were made (see Figure 1, right). Hence, in order to implement a comparable out-of-sample approach, we replicated the structure of the analysis of DSBC five times, with each analysis adopting one of the five distinct lotteries in KTHDP as reference lottery $x_*$. For example, we applied the TWT method (Theorem 2) and estimated utilities with a standard microeconometric approach using only the four binary choices involving option $[\$500; 29\%]$ and then predicted the remaining six comparisons not involving this option. We did this for each possible lottery, and here we report the average predictive performance across the five different analyses.

Figure 9 illustrates the results for both datasets. First, in DSBC, the predictive performance of TWT, measured as the average out-of-sample proportion of correctly predicted choices, is 76.14% (median 77.13%, min 49.62%, max 100%).

[13] Appendix C contains a further prediction exercise based on mean absolute errors instead of choices. This can also be done with the TWT method under the stronger assumption of Fechner errors, which brings it closer to parametric models.

This is a reasonably-high performance.[14] In particular, the out-of-sample predictive performance of TWT is significantly higher than that of standard econometric techniques. Crucially, this observation holds independently of the particular utility function assumed (CRRA vs. CARA) and of assumptions on the shape of the noise (RUM vs. RPM). For DSBC (Figure 9, left) the average out-of-sample proportion of correctly predicted choices according to RUM (CRRA) is 62.78% (median 63.54%, min 16.67%, max 100%) which is significantly smaller than what TWT achieves (WRS, $N = 60$, $z = -4.800$, $p < 0.0001$). Significant differences are also found comparing TWT to the other microeconometric implementations, even accounting for multiple-test corrections (RUM-CARA 59.55%, $z = 7.697$, $p < 0.0001$; RPM-CRRA 55.24%, $z = 8.713$, $p < 0.0001$; RPM-CARA 56.81%, $z = 8.740$, $p < 0.0001$).[15]

The out-of-sample predictive performance of the RUM and RPM estimations is quite modest. One possible reason is noise. In DSBC, subjects were overall very inconsistent during the experiment, possibly due to the high number of trials, the presence of repetitions, and similarities among the lotteries. However, those very same reasons make the performance of TWT noteworthy.

The overall picture is very similar for KTHDP's dataset (Figure 9, right), in spite of the differences between the experiments. TWT achieves a reasonable predictive performance (mean 71.76%, median 77.83%, min 32.67%, max 100%) and outperforms standard econometric approaches. The average out-of-sample proportion of correctly predicted choices according to RUM-CRRA is 54.11% (median 53.33%, min 0.00%, max 100.00%) which is smaller than that of TWT (WRS, $N = 26$, $z = -3.087$, $p = 0.0013$). A similar result is obtained for the other comparisons (RUM-CARA 56.00%, $z = -3.188$ $p = 0.0008$; RPM-CRRA 59.11%, $z = 3.087$, $p = 0.0013$; RPM-CARA 54.56%, $z = 3.506$, $p = 0.0002$), and accounting for multiple-test corrections.

These results can be given different interpretations. On the one hand, the predictive performance we obtain is reasonable, especially if one uses the nonparametric methods of TWT. This raises hopes that violations of transitivity might not be an unsurmountable obstacle for the prediction of economic choices out of sample. On the other hand, the datasets we rely on, while of course noisy, contain considerable amounts of information, in terms of repeated choices and process data

[14] Alós-Ferrer, Fehr, and Netzer (2021) reports an out-of-sample predictive performance of 80.7% in a food choice study where the options were simple food snacks; Garagnani (2020) finds that the out-of-sample predictive performance of different risk elicitation tasks is below 68%.

[15] In Appendix B.2 we show that these results are robust across the different conditions and manipulations in DSBC (time pressure and lottery format).

(response times). One might thus give the more pessimistic interpretation that "three out of four" is certainly better than typically achieved in the literature, but it might be close to an upper bound, as long as predictions are based on transitive models.

# 6 Discussion

Are economic choices transitive? A long-standing discussion in economics has addressed this fundamental issue. A negative answer would have the power to shake the very foundations of applied microeconomic analysis, and empirical evidence to this effect has been, understandably, subjected to detailed scrutiny. In particular, evidence in favor of transitivity violations have been systematically criticized as deriving from behavioral noise.

In this paper we provide a new method which allows to reveal "preferences" even when they are not transitive, disentangling them from behavioral noise. the method is based on a (straightforward) generalization of recent preference revelation results which use both choice frequencies and response times. We apply this method to two distinct datasets and find conclusive evidence that, even when one fully disentangles behavioral noise from underlying preferences, transitivity violations are reduced but do not disappear. In this sense, transitivity violations are not a mere artifact of the analysis or a consequence of behavioral noise, but rather an actual feature of human behavior.

These violations obviously hinder the ability of standard econometric techniques to predict choices, and hence have serious implications for positive and welfare economics. However, we also illustrate that non-parametric methods using response times still retain a reasonable predictive performance (of around 75% of correctly-predicted choices out of sample in our datasets) in spite of the incorrect transitivity assumption.

Although the latter results can be seen as good news, the now-undeniable existence of transitivity violations strongly suggests that descriptive models of choice assuming transitivity will eventually hit a ceiling in terms of their applicability. Thus, theories that dispense with the transitivity assumption might ultimately be needed to make progress beyond that ceiling.

# References

Alós-Ferrer, Carlos, Ernst Fehr, and Nick Netzer. 2021. "Time Will Tell: Recovering Preferences when Choices are Noisy." *Journal of Political Economy*

129 (6):1828–1877.

Alós-Ferrer, Carlos and Michele Garagnani. 2022a. "Strength of Preference and Decisions Under Risk." *Journal of Risk and Uncertainty* 64 (3):309–329.

———. 2022b. "The Gradual Nature of Economic Errors." *Journal of Economic Behavior and Organization* 200:55–66.

Anderson, Simon P., Jacques-François Thisse, and André De Palma. 1992. *Discrete Choice Theory of Product Differentiation*. Cambridge, MA: MIT Press.

Apesteguía, José and Miguel A. Ballester. 2018. "Monotone Stochastic Choice Models: The Case of Risk and Time Preferences." *Journal of Political Economy* 126 (1):74–106.

Baldassi, Carlo, Simone Cerreia-Vioglio, Fabio Maccheroni, and Massimo Marinacci. 2020. "A Behavioral Characterization of the Drift Diffusion Model and its Multi-Alternative Extension to Choice under Time Pressure." *Management Science* 66 (11):5075–5093.

Birnbaum, Michael H. 2013. "True-and-Error Models Violate Independence and yet They Are Testable." *Judgment and Decision making* 8 (6):717–737.

———. 2020. "Reanalysis of Butler and Pogrebna (2018) Using True and Error Model." *Judgment and Decision Making* 15 (6):1044–1051.

Birnbaum, Michael H. and Roman J. Gutierrez. 2007. "Testing for Intransitivity of Preferences Predicted by a Lexicographic Semi-Order." *Organizational Behavior and Human Decision Processes* 104 (1):96–112.

Birnbaum, Michael H., Jamie N. Patton, and Melissa K. Lott. 1999. "Evidence Against Rank-Dependent Utility Theories: Tests of Cumulative Independence, Interval Independence, Stochastic Dominance, and Transitivity." *Organizational Behavior and Human Decision Processes* 77 (1):44–83.

Birnbaum, Michael H. and Ulrich Schmidt. 2008. "An Experimental Investigation of Violations of Transitivity in Choice Under Uncertainty." *Journal of Risk and Uncertainty* 37 (1):77–91.

———. 2010. "Testing Transitivity in Choice Under Risk." *Theory and Decision* 69 (4):599–614.

Birnbaum, Michael H. and Lucy Wan. 2020. "MARTER: Markov True and Error Model of Drifting Parameters." *Judgment & Decision Making* 15 (1):47–73.

Bleichrodt, Han and Ulrich Schmidt. 2002. "A Context-Dependent Model of the Gambling Effect." *Management Science* 48 (6):802–812.

Block, Henry D. and Jacob Marschak. 1960. "Random Orderings and Stochastic Theories of Responses." In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, edited by Ingram Olkin. Stanford: Stanford University Press, 97–132.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice under Risk." *Quarterly Journal of Economics* 127 (3):1243–1285.

———. 2013. "Salience and Consumer Choice." *Journal of Political Economy* 121 (5):803–843.

Budescu, David V. and Wendy Weiss. 1987. "Reflection of Transitive and Intransitive Preferences: A Test of Prospect Theory." *Organizational Behavior and Human Decision Processes* 39 (2):184–202.

Butler, David. 2020. "Intransitive Preferences or Choice Errors? A Reply to Birnbaum." *Judgment and Decision Making* 15 (6):1052–1053.

Butler, David and Ganna Pogrebna. 2018. "Predictably Intransitive Preferences." *Judgment and Decision Making* 13 (3):217–236.

Cattell, James McKeen. 1893. "On Errors of Observation." *The American Journal of Psychology* 5 (3):285–293.

———. 1902. "The Time of Perception as a Measure of Differences in Intensity." *Philosophische Studien* 19:63–68.

Cavagnaro, Daniel R. and Clintin P. Davis-Stober. 2014. "Transitive in Our Preferences, but Transitive in Different Ways: An Analysis of Choice Variability." *Decision* 1 (2):102–122.

Chabris, Christopher F., Carrie L. Morris, Dmitry Taubinsky, David Laibson, and Jonathon P. Schuldt. 2009. "The Allocation of Time in Decision-Making." *Journal of the European Economic Association* 7 (2-3):628–637.

Clithero, John A. 2018. "Improving Out-of-Sample Predictions Using Response Times and a Model of the Decision Process." *Journal of Economic Behavior and Organization* 148:344–375.

Conte, Anna, John D. Hey, and Peter G. Moffatt. 2011. "Mixture Models of Choice Under Risk." *Journal of Econometrics* 162 (1):79–88.

Dashiell, John F. 1937. "Affective Value-Distances as a Determinant of Aesthetic Judgment-Times." *American Journal of Psychology* 50:57–67.

Davis-Stober, Clintin P., Nicholas Brown, and Daniel R. Cavagnaro. 2015. "Individual Differences in the Algebraic Structure of Preferences." *Journal of Mathematical Psychology* 66:70–82.

Davis-Stober, Clintin P., Denis M. McCarthy, Daniel R Cavagnaro, Mason Price, Nicholas Brown, and Sanghyuk Park. 2019. "Is Cognitive Impairment Related to Violations of Rationality? A Laboratory Alcohol Intoxication Study Testing Transitivity of Preference." *Decision* 6 (2):134–144.

Dehaene, Stanislas, Emmanuel Dupoux, and Jacques Mehler. 1990. "Is Numerical Comparison Digital? Analogical and Symbolic Effects in Two-Digit Number Comparison." *Journal of Experimental Psychology: Human Perception and Performance* 16 (3):626–641.

Diecidue, Enrico and Jeeva Somasundaram. 2017. "Regret Theory: A New Foundation." *Journal of Economic Theory* 172:88–119.

Fishburn, Peter C. 1971. "A Study of Lexicographic Expected Utility." *Management Science* 17 (11):672–678.

———. 1982. "Nontransitive Measurable Utility." *Journal of Mathematical Psychology* 26 (1):31–67.

———. 1984a. "Dominance in SSB Utility Theory." *Journal of Economic Theory* 34 (1):130–148.

———. 1984b. "SSB Utility Theory: An Economic Perspective." *Mathematical Social Sciences* 8 (1):63–94.

———. 1984c. "SSB Utility Theory and Decision-Making Under Uncertainty." *Mathematical Social Sciences* 8 (3):253–285.

———. 1986. "Ordered Preference Differences Without Ordered Preferences." *Synthese* 67 (2):361–368.

———. 1988. *Nonlinear Preference and Utility Theory*. Baltimore, Maryland: Johns Hopkins University Press.

———. 1991. "Nontransitive Preferences in Decision Theory." *Journal of Risk and Uncertainty* 4 (2):113–134.

———. 1998. "Stochastic Utility." In *Handbook of Utility Theory*, vol. 1: Principles, edited by Salvador Barberà, Peter J. Hammond, and Christian Seidl, chap. 7. Kluwer Academic Publishers, 273–319.

Fudenberg, Drew, Philipp Strack, and Tomasz Strzalecki. 2018. "Speed, Accuracy, and the Optimal Timing of Choices." *American Economic Review* 108 (12):3651–3684.

Garagnani, Michele. 2020. "The Predictive Power of Risk Elicitation Tasks." Working Paper, University of Zurich.

González-Vallejo, Claudia. 2002. "Making Trade-Offs: A Probabilistic and Context-Sensitive Model of Choice Behavior." *Psychological Review* 109 (1):137–155.

Grether, David M. and Charles R. Plott. 1979. "Theory of Choice and the Preference Reversal Phenomenon." *American Economic Review* 69 (4):623–638.

Hatz, Laura E., Sanghyuk Park, Kayleigh N. McCarty, Denis M. McCarthy, and Clintin P. Davis-Stober. 2020. "Young Adults Make Rational Sexual Decisions." *Psychological Science* 31 (8):944–956.

Hausner, Melvin. 1954. "Multidimensional Utilities." *Decision Processes* :167–180.

Herweg, Fabian and Daniel Müller. 2019. "Regret Theory and Salience Theory: Total Strangers, Distant Relatives or Close Cousins?"

———. 2021. "A Comparison of Regret Theory and Salience Theory for Decisions Under Risk." *Journal of Economic Theory* 193:105226.

Humphrey, Steven. 2001. "Non-transitive Choice: Event-Splitting Effects or Framing Effects?" *Economica* 68 (269):77–96.

Iverson, Geoffrey and Jean-Claude Falmagne. 1985. "Statistical Issues in Measurement." *Mathematical Social Sciences* 10 (2):131–153.

Kalenscher, Tobias, Philippe N. Tobler, Willem Huijbers, Sander M. Daselaar, and Cyriel Pennartz. 2010. "Neural Signatures of Intransitive Preferences." *Frontiers in Human Neuroscience* 4 (49):1–14.

Klein, A. Stanley. 2001. "Measuring, Estimating, and Understanding the Psychometric Function: A Commentary." *Attention, Perception, & Psychophysics* 63 (8):1421–1455.

Kontek, Krzysztof. 2016. "A Critical Note on Salience Theory of Choice Under Risk." *Economics Letters* 149:168–171.

Krajbich, Ian, Björn Bartling, Todd Hare, and Ernst Fehr. 2015. "Rethinking Fast and Slow Based on a Critique of Reaction-Time Reverse Inference." *Nature Communications* 6 (7455):1–9.

Kreweras, G. 1961. "Sur une possibilité de rationaliser les intransitivités." In *La Décision, Colloques Internationaux du Centre National de la Recherche Scientifique*. Paris: Editions du Centre National de la Recherche Scientifique, 27–32.

Laming, Donald. 1985. "Some Principles of Sensory Analysis." *Psychological Review* 92 (4):462–485.

Lee, Leonard, On Amir, and Dan Ariely. 2009. "In Search of Homo Economicus: Cognitive Noise and the Role of Emotion in Preference Consistency." *Journal of Consumer Research* 36 (2):173–187.

Lee, Leonard, Michelle P. Lee, Marco Bertini, Gal Zauberman, and Dan Ariely. 2015. "Money, Time, and the Stability of Consumer Preferences." *Journal of Marketing Research* 52 (2):184–199.

Leland, Jonathan W. 1994. "Generalized Similarity Judgments: An Alternative Explanation for Choice Anomalies." *Journal of Risk and Uncertainty* 9 (2):151–172.

———. 1998. "Similarity Judgments in Choice Under Uncertainty: A Reinterpretation of the Predictions of Regret Theory." *Management Science* 44 (5):659–672.

Li, Zhihua and Graham Loomes. 2022. "Revisiting the Diagnosis of Intertemporal Preference Reversals." *Journal of Risk and Uncertainty* :1–23.

Lichtenstein, Sarah and Paul Slovic. 1971. "Reversals of Preference Between Bids and Choices in Gambling Decisions." *Journal of Experimental Psychology* 89 (1):46–55.

Lindman, Harold R. and James Lyons. 1978. "Stimulus Complexity and Choice Inconsistency Among Gambles." *Organizational Behavior and Human Performance* 21 (2):146–159.

Loomes, Graham, Chris Starmer, and Robert Sugden. 1989. "Preference Reversal: Information-Processing Effect or Rational Non-Transitive Choice?" *Economic Journal* 99 (395):140–151.

———. 1991. "Observing Violations of Transitivity by Experimental Methods." *Econometrica* 59 (2):425–439.

Loomes, Graham and Robert Sugden. 1982. "Regret Theory: An Alternative Theory of Rational Choice Under Uncertainty." *Economic Journal* 92 (368):805–824.

———. 1987. "Some Implications of a More General Form of Regret Theory." *Journal of Economic Theory* 41 (2):270–287.

———. 1998. "Testing Different Stochastic Specifications of Risky Choice." *Economica* 65 (260):581–598.

Luce, R. Duncan. 1959. *Individual Choice Behavior: A Theoretical Analysis*. New York: Wiley.

———. 2000. *Utility of Gains and Losses: Measurement-Theoretical and Experimental Approaches*. New York, NY: Psychology Press.

Marley, Anthony A.J. and R. Duncan Luce. 2005. "Independence Properties vis-à-vis Several Utility Representations." *Theory and Decision* 58 (1):77–143.

Marschak, Jacob. 1960. "Binary Choice Constraints on Random Utility Indicators." In *Stanford Symposium on Mathematical Methods in the Social Sciences*, edited by Kenneth J. Arrow. Stanford, CA: Stanford University Press, 312–329.

May, Kenneth O. 1954. "Intransitivity, Utility, and the Aggregation of Preference Patterns." *Econometrica* 22 (1):1–13.

McFadden, Daniel L. 1974. "Conditional Logit Analysis of Qualitative Choice Behavior." In *Frontiers in Econometrics*, edited by P. Zarembka. New York: Academic Press, 105–142.

———. 2001. "Economic Choices." *American Economic Review* 91 (3):351–378.

Moffatt, Peter G. 2005. "Stochastic Choice and the Allocation of Cognitive Effort." *Experimental Economics* 8 (4):369–388.

Montgomery, Henry. 1977. "A Study of Intransitive Preferences Using a Think Aloud Procedure." In *Decision Making and Change in Human Affairs*. 347–362.

Mosteller, Frederick and Philip Nogee. 1951. "An Experimental Measurement of Utility." *Journal of Political Economy* 59:371–404.

Moyer, Robert S. and Richard H. Bayer. 1976. "Mental Comparison and the Symbolic Distance Effect." *Cognitive Psychology* 8 (2):228–246.

Moyer, Robert S. and Thomas K. Landauer. 1967. "Time Required for Judgements of Numerical Inequality." *Nature* 215 (5109):1519–1520.

Müller-Trede, Johannes, Shlomi Sher, and Craig R. M. McKenzie. 2015. "Transitivity in Context: A Rational Analysis of Intransitive Choice and Context-Sensitive Preference." *Decision* 2 (4):280–305.

Park, Sanghyuk, Clintin P. Davis-Stober, Hope K Snyder, William Messner, and Michel Regenwetter. 2019. "Cognitive Aging and Tests of Rationality." *The Spanish Journal of Psychology* 22 (E57):1–26.

Ranyard, Rob, Henry Montgomery, Emmanouil Konstantinidis, and Andrea Louise Taylor. 2020. "Intransitivity and Transitivity of Preferences: Dimensional Processing in Decision Making." *Decision* 7 (4):287–313.

Ratcliff, Roger. 1978. "A Theory of Memory Retrieval." *Psychological Review* 85:59–108.

Regenwetter, Michel, Jason Dana, and Clintin P. Davis-Stober. 2010. "Testing Transitivity of Preferences on Two-Alternative Forced Choice Data." *Frontiers in Psychology* 1 (148):1–15.

———. 2011. "Transitivity of Preferences." *Psychological Review* 118 (1):42–56.

Savage, Leonard J. 1954. *The Foundations of Statistics*. New York: John Wiley & Sons.

Savage Jr., Richard P. 1994. "The Paradox of Nontransitive Dice." *American Mathematical Monthly* 101 (5):429–436.

Schmidt, Ulrich and Michael Stolpe. 2011. "Transitivity in Health Utility Measurement: An Experimental Analysis." *Health Economics Review* 1 (1):1–12.

Shafer, Wayne J. 1974. "The Nontransitive Consumer." *Econometrica* 42:913–919.

Shafir, Sharoni. 1994. "Intransitivity of Preferences in Honey Bees: Support for Comparative Evaluation of Foraging Options." *Animal Behaviour* 48 (1):55–67.

Sopher, Barry and Gary Gigliotti. 1993. "Intransitive Cycles: Rational Choice or Random Error? An Answer Based on Estimation of Error Rates with Experimental Data." *Theory and Decision* 35 (3):311–336.

Starmer, Chris and Robert Sugden. 1998. "Testing Alternative Explanations of Cyclical Choices." *Economica* 65 (259):347–361.

Thurstone, Louis L. 1927. "A Law of Comparative Judgement." *Psychological Review* 34:273–286.

Tversky, Amos. 1969. "Intransitivity of Preferences." *Psychological Review* 76:31–48.

Tversky, Amos and Daniel Kahneman. 1992. "Advances in Prospect Theory: Cumulative Representation of Uncertainty." *Journal of Risk and Uncertainty* 5 (4):297–323.

Tversky, Amos, Paul Slovic, and Daniel Kahneman. 1990. "The Causes of Preference Reversal." *American Economic Review* 80 (1):204–217.

Tversky, Amos and Richard H. Thaler. 1990. "Anomalies: Preference Reversals." *Journal of Economic Perspectives* 4 (2):201–211.

Waite, Thomas A. 2001. "Intransitive Preferences in Hoarding Gray Jays (Perisoreus Canadensis)." *Behavioral Ecology and Sociobiology* 50 (2):116–121.

Wichmann, A. Felix and N. Jeremy Hill. 2001. "The Psychometric Function: I. Fitting, Sampling, and Goodness of Fit." *Attention, Perception, & Psychophysics* 63 (8):1293–1313.

# (ONLINE) APPENDIX

## A Revealing Preferences and Estimating Risk

Theorem 2 of TWT allows to make out-of-sample predictions assuming symmetric noise (but no further assumptions, and no structural restrictions) as long as the dataset includes repeated decisions between each individual alternative and a fixed, reference one. This is the exact structure of the DSBC dataset, and can be reproduced in the KTHDP one since all comparisons are made there. For the implementation of the out-of-sample predictions following the TWT method, we followed Alós-Ferrer, Fehr, and Netzer (2021) and refer the reader to that article for further details.

We detail now our predictive analyses following standard microeconometric models. In order to compare the performance of these prediction exercises and TWT, we estimate each subject's risk attitude from the appropriate set of binary lottery choices following a standard maximum likelihood estimation. All trials used for the estimation involved binary choices between lotteries of the form $A = (p, x)$ and $B = (q, y)$, where A pays $x$ with probability $p$ and B pays $y$ with probability $q$, and 0 otherwise (see Figure 1 in the main text for the actual lotteries). We index the trials in the experiments by $t = 1, \ldots, N$. That is, in trial $t$ subjects face the choice between $A_t = (p_t, x_t)$ and $B_t = (q_t, y_t)$. In the main analysis we assume two different utility functions. The first is a normalized constant absolute risk aversion (CARA) function as in Conte, Hey, and Moffatt (2011), given by

$$u(x \mid r) = \begin{cases} \frac{1-e^{-rx}}{1-e^{-rx_{\max}}}, & \text{if } r \neq 0 \\ \frac{x}{x_{\max}}, & \text{if } r = 0, \end{cases}$$

where $x_{\max} = \max\{x_1, \ldots, x_N, y_1, \ldots, y_N\}$ is the maximum outcome across all $N$ lottery pairs (trials). The normalization ensures that $u(x \mid r)$ is increasing also for negative values of $r$ (indicating risk-seeking behavior). The second utility form is a constant relative risk aversion (CRRA) function $u(x \mid r) = x^r$. Under the assumption of Expected Utility maximization, subject $i$ with utility function (CARA) $u(x \mid r_i)$ chooses $A_t$ over $B_t$ if the difference in expected utilities is positive, that is,

$$(1) \qquad \nabla_t(r_i) := p_t u(x_t \mid r_i) - q_t u(y_t \mid r_i) = \frac{p_t(1 - e^{-r_i x_t}) - q_t(1 - e^{-r_i y_t})}{1 - e^{-r_i x_{\max}}} > 0,$$

and analogously for the CRRA case. The second element of the model is a noise term. For these models, there are two standard approaches to noise in the literature: The Fechner or Random Utility Model (RUM) and the Random Preference Model (RPM). RUM assumes that each subject is characterized by a risk parameter $r_i$ that is fixed across trials, but utility is affected by an additive utility-noise term with a fixed distribution, e.g. normal. In contrast, RPM assumes that a subject's risk parameter varies randomly between trials but is drawn from a certain distribution. We present the results of our analyses using both approaches.

Following the RUM approach, we add an error term $\varepsilon_{it} \sim N(0, \sigma_i^2)$ with $\sigma_i^2 > 0$ to (1). That is, the lottery $A_t$ is chosen if

$$\nabla_t(r_i) + \varepsilon_{it} > 0.$$

Define the binary choice indicator for trial $t$

$$\gamma_{it} = \begin{cases} 1 & \text{if } A_t \text{ chosen by subject } i \\ -1 & \text{if } B_t \text{ chosen by subject } i. \end{cases}$$

Then the probability of a choice conditional on the risk-parameter $r_i$ is given by

$$p(\gamma_{it} \mid r_i) = P\left(\gamma_{it}\nabla_t(r_i) > \gamma_{it}(-\varepsilon_{it})\right) = P\left(\gamma_{it}\frac{\nabla_t(r_i)}{\sigma} > \gamma_{it}\frac{-\varepsilon_{it}}{\sigma}\right) = \Phi\left(\gamma_{it}\frac{\nabla_t(r_i)}{\sigma}\right)$$

where $\Phi$ is the standard normal cumulative distribution function. We used maximum likelihood to estimate individual risk attitudes based on the conditional probability above. The procedure delivers estimates of the individual risk attitude $r_i$ and the variance of the (normally-distributed) error term, $\sigma_i^2$.

For the RPM estimation, we use the same subsets of choices for the predictive analyses as in the RUM case, and again we consider both CARA and CRRA utility functions. Additionally, we re-arrange the dataset in such a way that $A_t$ is always the safer of the two lotteries, that is, $p > q$ (no dominated lotteries are considered in the analysis). In contrast to the RUM approach, the RPM assumes that a subject's risk parameter is not fixed across trials but varies randomly between trials. Specifically, we assume that subject $i$'s risk parameter in trial $t$ is distributed according to $r_{it} \sim N(m_i, \sigma_i^2)$ where $m_i$ is subject $i$'s mean risk attitude. Assuming Expected Utility maximization, in this setup subject $i$ with utility function $u_i$ chooses $A_t$ over $B_t$ if and only if

$$\Delta_t(r_{it}) = \frac{p_t(1 - e^{-r_{it}x_t}) - q_t(1 - e^{-r_{it}y_t})}{1 - e^{-r_{it}x_{\max}}} > 0.$$

Let $r_t^*$ be the risk parameter that would make a subject exactly indifferent between the two lotteries in task $t$, that is, $\Delta_t(r_t^*) = 0$. Since $A_t$ is always the safer lottery, we obtain the following equivalence

$$\Delta_t(r_{it}) > 0 \quad \Leftrightarrow \quad r_{it} > r_t^*.$$

Again using $\gamma_{it} \in \{1, -1\}$ as a binary indicator that $A_t$ is chosen by subject $i$ in trial $t$, the probability of a choice conditional on a subject's mean risk attitude $m_i$ is given by

$$p(\gamma_{it}|m_i) = p(\gamma_{it}r_{it} > \gamma_{it}r_t^*|m_i) = P\left(\gamma_{it}\frac{r_{it} - m_i}{\sigma_i} > \gamma_{it}\frac{r_t^* - m_i}{\sigma_i}\right) = \Phi\left(\gamma_{it}\frac{m_i - r_t^*}{\sigma_i}\right)$$

where $\Phi$ is the standard normal cumulative distribution function. Again, we relied on maximum likelihood estimation based on the conditional probability above. This delivers estimates of the mean $m_i$ and the variance $\sigma_i^2$.

# B Robustness Analysis: Time Pressure and Lottery Formats

In DSBC two *within-subject* treatments were implemented, time pressure vs. no time pressure and pie vs. bar lottery format. We can hence investigate the possible influence of these manipulations on our results.

## B.1 Transitivity Violations

We start with preference revelation and transitivity violations. Comparing revealed preferences over binary choices, there are no statistical differences between time pressure and its absence (56.13% vs. 57.16%; WRS $N = 60$, $z = -0.942$, $p = 0.3505$). However, we observe that using the bar representation is associated with a higher proportion of revealed preferences (59.87%) compared to the pie representation (53.84%; WRS $N = 60$, $z = 3.872$, $p < 0.001$).

Comparing overall proportions of transitivity violations, again there are no statistically significant differences between time pressure and its absence (20.32% vs. 21.21%; WRS $N = 60$, $z = -0.578$, $p = 0.5671$). A similar result is obtained when we consider RTV (19.03% vs. 19.15%; WRS $N = 60$, $z = -0.129$, $p = 0.9011$). However, pie representations lead to a larger proportion of transitivity violations compared the bar representations, although the comparison misses significance at the 5% level (21.35% vs. 20.19%; WRS $N = 60$, $z = 1.716$, $p = 0.0866$). There are no significant differences when we consider RTV (18.52% vs. 20.04%; WRS $N = 60$, $z = -0.648$, $p = 0.5222$).

## B.2 Prediction Exercise

In the main text we show that the out-of-sample predictive performance of TWT (Theorem 2) is systematically higher than that of RUMs or RPMs. In this subsection we report on some illustrative robustness analyses for the different conditions in the DSBC dataset.

Figure B.1 shows that TWT outperforms RUM in all different conditions implemented in the design of Davis-Stober, Brown, and Cavagnaro (2015) (time pressure and lottery format). The figure displays data for the RUM-CARA case, but results are robust to other utility or noise specifications. The proportion of correctly predicted choices is higher in TWT for time pressure, both with pie representations (73.79% vs. 56.52%; WRS, $N = 60$, $z = 3.539$, $p = 0.0004$) and bar representations (75.31% vs. 63.06%; WRS, $N = 60$, $z = 2.918$, $p = 0.0035$), and also for the absence of time pressure, again both for pie representations (76.97% vs. 59.58%; WRS, $N = 60$, $z = 3.977$, $p < 0.0001$), and bar representations (78.48% vs. 57.22%; WRS, $N = 60$, $z = 4.199$, $p < 0.0001$).

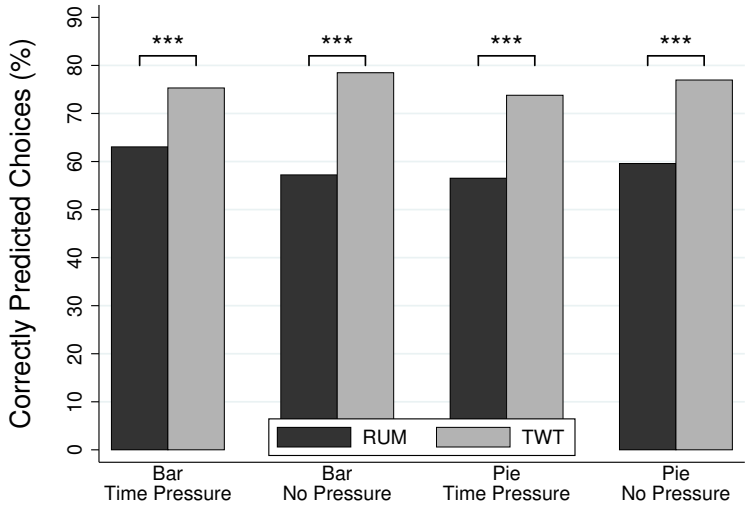


Figure B.1: Proportion of out-of-sample correctly predicted choices for Davis-Stober, Brown, and Cavagnaro (2015) across the four different conditions implemented in the experiment.

# C Predictive Performance Assuming Fechner Errors

The standard microeconometric approach that we used in the main text involves a probit model, i.e. normally-distributed, hence Fechnerian errors. If one is willing to assume Fechnerian errors (as in any logit or probit model) then Theorem 3 of Alós-Ferrer, Fehr, and Netzer (2021) provides a method to predict the proportion of choices and not just the binary relation, without assuming a specific functional form for utilities or a specific functional shape of the noise term beyond the fact that noise must be Fechnerian.[16] To quantify the predictive performance of the analyses, we use the mean absolute error as in Alós-Ferrer, Fehr, and Netzer (2021) and Clithero (2018). This measure calculates the individual average distance between predicted and observed choice frequencies (results are similar using alternative measures as, e.g., the squared root of the sum of the squared differences).

This result is as follows. Within the class of Fechnerian RUM-CFs, a rationalizable SCF-RT *predicts choice probability* $\bar{p}(x,y)$ for a non-observed choice $(x,y) \in C \setminus D$ if all RUM-CFs in the class that rationalize it satisfy $\text{Prob}[\tilde{v}(x,y) > 0] = \bar{p}(x,y)$.

[16]Noise in a RNM-CF is Fechnerian if, for each $(x,y) \in C$ and all $v \in \mathbb{R}$, $g(x,y)(v) = g(v - v(x,y))$, where $G$ is a common density $g$ with full support such that $g(\delta) = g(-\delta) > 0$ for all $\delta \geq 0$.

**Theorem 3** (Theorem 3, Alós-Ferrer, Fehr, and Netzer, 2021)**.** *Let $(x, y) \in C \setminus D$ and $x_* \in X$ with $(x, x_*), (y, x_*) \in D$. Within the class of Fechnerian RUM-CFs, a rationalizable SCF-RT predicts the choice probability*

$$\bar{p}(x,y) = \begin{cases} p(x,z)F(x,z)(\theta(y,z)) & \text{if } p(y,z) > 1/2, \\ p(x,z) & \text{if } p(y,z) = 1/2, \\ 1 - p(z,x)F(z,x)(\theta(z,y)) & \text{if } p(y,z) < 1/2. \end{cases}$$

For the out-of-sample prediction, we follow the same approach as in the main text. That is, we again rely on the particular structure of DSBC's dataset (see Figure 8) to predict choice frequencies in decisions not involving $x_*$ after estimating preferences and noise parameters from the decisions involving $x_*$. For KTHDP, again we average the five possible out-of-sample exercises (taking each distinct lottery in the dataset as the reference).

The microeconometric estimation used in the main text obviously also allows to predict choice frequencies (instead of deterministic binary choices), making the prediction comparable with that of Theorem 3 above. For this purpose, we use the estimates described in Appendix A to compute the predicted choice frequencies in the corresponding RUM or RPM models. That is, instead of predicting an alternative for each binary choice, we use the estimated risk attitude and noise variance (for the RUM case) or the estimated mean and variance of the individual risk attitudes (for the RPM case) to predict choice frequencies. In particular, for the RUM approach we predict that a subject will choose option $A_t$ over $B_t$ at trial $t$ with probability

$$p = \Phi\left(\frac{\nabla_t(r_i)}{\sigma}\right)$$

Similarly, for the RPM approach we predict that a subject will choose the option $A_t$ over $B_t$ at trial $t$ with probability

$$p = \Phi\left(\frac{m_i - r_t^*}{\sigma_i}\right).$$

Again, to measure the accuracy of our prediction, we compute the mean absolute error between the observed choice proportions and the predicted proportions.

The results are shown in Figure C.1 (recall that a good performance corresponds to a small mean absolute error). For each individual, we compute the mean absolute error. For DSBC, the average mean absolute error across individuals for the TWT method is 0.2120.[17] This outperforms the results when using a RUM estimation with CARA utility functions (0.3316; WRS, $N = 60$, $z = -5.926$, $p < 0.0001$) or an RPM approach with either CRRA (0.3787; WRS, $N = 60$, $z = -6.618$, $p < 0.0001$) or CARA utilities (0.3765; WRS, $N = 60$, $z = -6.530$, $p < 0.0001$). However, the performance of a RUM estimation using CRRA functions is better than that of TWT for this dataset (0.1386; WRS, $N = 60$, $z = 5.875$, $p < 0.0001$).

[17] Alós-Ferrer, Fehr, and Netzer (2021) obtain 0.237 with the food choice data of Clithero (2018). The latter obtains 0.209 using a parametric drift-diffusion model approach.

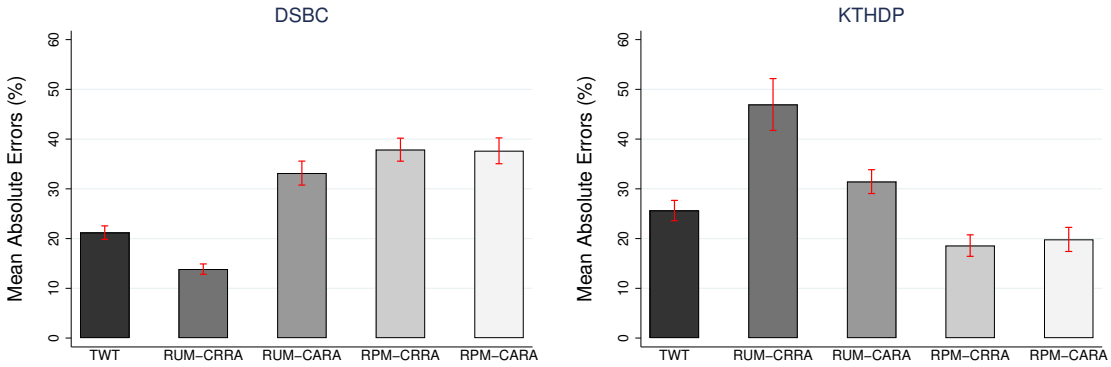


Figure C.1: Mean absolute errors for RUM/RPM and TWT for Davis-Stober, Brown, and Cavagnaro (2015) (on the left) and Kalenscher et al. (2010) (on the right) across different utility functions (CRRA vs. CARA). 95% confidence intervals are represented in red.

A qualitatively similar result is obtained for KTHDP's dataset (Figure C.1, right). The TWT approach achieves an average mean absolute error of 0.2564, which outperforms RUM estimations with either CRRA (0.4695; WRS, $N = 29$, $z = 4.076$, $p < 0.0001$) or CARA utilities (0.3145; WRS, $N = 24$, $z = 2.286$, $p = 0.0211$). However, RPM estimations outperform TWT for this dataset, both with CRRA (0.1859 WRS, $N = 30$, $z = -2.910$, $p = 0.0028$) and with CARA utilities (0.1982 WRS, $N = 30$, $z = -2.088$, $p = 0.0364$), although these results do not reach significance if adjusting for multiple testing.

The mixed results for the application of TWT's Theorem 3 might simply reflect the dangers of additional, possibly-unwarranted assumptions used in estimation procedures. Theorem 3 in the TWT method assumes Fechner errors (although not a specific functional shape), an assumption that might be less warranted than simply symmetric errors as in Theorem 2. Fechner errors reduce the conceptual distance between the TWT method and the RUM approach (with normally-distributed errors) or RPM analyses (with normally-distributed risk attitudes). While without this assumption the TWT method significantly outperformed the other, parametric approaches, adopting this assumption leads to mixed results, which are also inconsistent across datasets.