Dynamiken bakom bevarandet av transitiva val och preferenser: Transitivitet i kontexten av Choice Blindness

The Dynamics of Preserving Transitive Choice and Preference: Transitivity in the context of Choice Blindness

Daniel Andreas Zander

Handledare / Supervisors

Petter Johansson
Gabriel Vogel

KOGM20

7th June 2023

Masteruppsats i kognitionsvetenskap
Avdelningen för kognitionsvetenskap
Filosofiska institutionen
Lunds universitet

Master's thesis (2 years) in Cognitive Science
Lund University Cognitive Science
Department of Philosophy
Lund University

The Dynamics of Preserving Transitive Choice and Preference:
Transitivity in the context of Choice Blindness

Daniel Andreas Zander
da8183za-s@student.lu.se

7th June 2023

In most normative and prescriptive decision theories, transitivity is a presumed property of any preference relation, stating that for any choice alternatives x, y, z and preference relation R, if xRy and yRz, then xRz. However, what is normatively expected of an ideally rational agent may still be descriptively false under certain contexts and empirically motivated decision models. One such possible context is that of Choice Blindness manipulation. Current investigations attempts a purposeful elicitation of intransitivity by utilizing preference change effects following choice blindness manipulation. To the extent that this is possible, the thesis seeks to discern possible cognitive mechanisms that contribute to the defence or preservation of transitivity in choice and preference. Initial results indicate no effect of choice blindness manipulation on the rates of intransitive preference. Further analysis demonstrates different patterns of behavior which are most readily interpreted as attempts at defending against intransitive or otherwise preserving already transitive choice structures.

1 Introduction

Imagine you're standing in the grocery store investigating the ripeness of three bananas x, y and z. Two different pair-wise comparisons allows you to judge that x is *more ripe* than y and y more ripe than z. Before comparing x and z you come to realize that "more ripe than" is a *transitive* relation on the domain of bananas, and so you proudly infer that x must be more ripe than z! Based on a confluence of factors (of which ripeness is but one) you likewise judge your *preference* for x to be greater than y and your preference for y to be greater than z. Given the facts of the situation you can now ask two questions about your pair-wise preference of x and z: (1) am I *rationally* obliged to prefer x over z? (2) Do I and people in relevantly similar situations *tend* to prefer x over z? (1) asks a normative and (2) a descriptive question with regards to the *transitive* nature of human preference.

Although obviously interrelated, an affirmative answer to one of these question does not force the issue with regards to the other. In classical (real-valued) quantitative attempts to model preference, transitivity is both a formal constraint and a presumed approximation to preference-driven choice behavior (Bernoulli, 1738; Friedman & Savage, 1952; Von Neumann & Morgenstern, 1947). An influential argument for the normative claim employs a *pragmatic* justification, showing how *in*transitive individuals are subject to exploitative "money pumps" (Bar-Hillel & Margalit, 1988; Block, Barnett et al., 2012). Be that as it may, going back to Tversky (1969) there has been plenty of decision-contexts in which people's preferences seemingly violate transitivity (e.g. Brandstätter, Gigerenzer & Hertwig, 2006; Butler & Pogrebna, 2018; Kivetz & Simonson, 2000; Loomes & Sugden, 1987). However, it is notoriously hard to show that these are *true* cases of intransitive preference as opposed to the result of inherent variability in choice data (Regenwetter, Dana & Davis-Stober, 2011). Properly testing the empirical validity of transitivity presupposes an appropriate translation of what is essentially a deterministic algebraic axiom, to a statistical model which accommodates the random sampling process of empirical data (Luce, 1995, 1997).

This thesis aspires to contribute to the existing literature with a study on binary choice and preference in the context of *Choice Blindness*. This involves the use of a common two-alternative forced-choice task (2AFC) with two repeated pair-wise choices per comparison, where the object of analysis is choice (and by extension preference) defined over triples of choice alternatives. False feedback is used on certain pair-wise comparisons in order to elicit intransitivity in people's preferences. Although common hypothesis testing and modelling will be used, the choice blindness phenomena allows for certain novel indirect ways at probing the cognitive alignment with transitivity. As such, the thesis is driven by an indirect question as to the potential existence of a cognitive *barrier* against intransitive choice and preference. The research question(s) can be formulated thus:

> **(Q)** Is cognition equipped with mechanisms for preserving transitive or defending against intransitive choice and preference?
>
> ⋮
>
> **(q)** Can choice blindness manipulation reliably induce intransitive preferences?

An affirmative answer to **(q)** is suggestive of a negative answer to **(Q)**, and vice versa. Answering **(Q)** in the affirmative involves statements of hypothesized transitivity preserving behavior. That is, patterns of behavior which effectively counteracts the intrusive promotion of *in*transitivity in the choice blindness context. By adopting this partially indirect approach to studying transitivity, the ambition is to gain novel insights and potentially sidestep some of the methodological challenges encountered in previous research efforts.

The Choice Blindness Phenomena

Change blindness demonstrates our failure to, across sensory modalities, notice quite ostensible changes in our immediate perceptual environment (Auvray, Gallace, Hartcher-O'Brien, Tan & Spence, 2008; Dickerson & Gaston, 2014; Simons & Rensink, 2005). The phenomena of *choice* blindness extends this observation to the realm of *intention* and *action*, where particular changes to one's perceptual field is of considerable

subjective importance (Johansson, Hall & Sikström, 2008). In their original paper, Johansson, Hall, Sikstrom and Olsson (2005) asked participants to make pair-wise choices between faces based on attractiveness. On some trials they then gave participants *false feedback* in relation to their prior intention, showing and asking participants to give reasons for selecting a face *opposite* of their actual choosing. In the great majority of manipulated trials participants failed to notice the mismatch between prior intention and the outcome of their choice; even going on to report *confabulated* reasons for their behavior.

In a subsequent study, Johansson, Hall, Tärning, Sikström and Chater (2014) incorporated a second round of choice into their experimental procedure, which included all the face-pairs from the first round. This allowed for a measure of *choice consistency* between the two rounds of choice, i.e. the extent to which the same face in each pair was chosen the second time around. Results show that people are very consistent with their first round choice in non-manipulated conditions, but that this rate drops significantly for manipulated face-pairs. Choice blindness manipulation (henceforth, CB-manipulation) can therefore be seen to change peoples preferences in line with the previously non-chosen alternative. This effect has since been replicated in multiple studies (Izuma et al., 2015; Luo & Yu, 2017; Strandberg, Sivén, Hall, Johansson & Pärnamets, 2018; Taya, Gupta, Farber & Mullette-Gillman, 2014).

This paper utilizes CB-manipulation and the observed preference change effect to study the transitivity of preference over triples of faces. An adapted version of the general procedure will be used to try and induce intransitive preference relations or otherwise investigate ways in which this is systematically avoided by participants. The specific nature of such avoidance or defense against intransitivity will be outlined in successive degrees of detail, beginning with a course formulation in the section titled *Hypotheses*.

Notation and definitions

Before moving on it will be useful to introduce notation and some accompanying formal definitions. In the study of transitivity, the paradigm of two-alternative forced choice goes back to Tversky (1969) and has since then been commonplace (Müller-Trede, Sher & McKenzie, 2015; Regenwetter, Dana & Davis-Stober, 2010; Regenwetter et al., 2011). It involves the *exclusive* selection between alternatives x_i and x_j in a pair-wise comparison $\{x_i, x_j\}$. This exclusiveness of choice assumes that underlying preference states can be modeled by a *strict* preference relation, denoted by $\succ$. Essentially, strict preference relations precludes a decision maker (DM) from being in states of *indifference* (denoted by $\sim$) between any two choice alternatives. From its *weak* order counterpart, $\succeq$, expressing "better than or equal in value to", $\succ$ can be defined as follows:

($\succ_{def}$) for any choice alternatives x_i and x_j,
$x_i \succ x_j \Leftrightarrow [x_i \succeq x_j \wedge \neg(x_i \preceq x_j)]$.

In words: x_i is strictly preferred to x_j if and only if x_i is of equal or greater value than x_j, but not the other way around.[1]Next, let $X = \{x_1, x_2, \ldots, x_i\}$ denote a finite collection of alternatives called a *choice set* (e.g. bananas, face pictures, cars). With the available notation, the *axiom of transitivity* states that:

(Transitivity) for any choice alternatives $x_i, x_j, x_k \in X$, $[x_i \succ x_j \wedge x_j \succ x_k] \Rightarrow x_i \succ x_k$.

A preference relation is *in*transitive when the above statement fails. Discussion of transitivity and its probabilistic formulations will follow in subsequent sections. To explicitly distinguish between choice and preference we make use of a binary *choice* relation $\succ^*$. $\psi \succ^* \phi$ means that an external agent has observed some DM make a *forced* choice of ψ on a comparison $\{\psi, \phi\}$ (inspired by Nishimura & Ok, 2018). What it means for $\succ^*$ to be transitive is of course no different from the case of $\succ$. $x_i \succ^* x_j \succ^* x_k$ is a compressed way to express a particular *pattern* of choice where $x_i \succ^* x_j, x_j \succ^* x_k$ and $x_i \succ^* x_k$. $(X, \succ^*)$ is a short-hand way of referencing a choice pattern defined on choice set X. One may think of the earlier expanded pattern (and any arbitrary one) as a set, $(X, \succ^*) = \{(x_i, x_j), (x_j, x_k), (x_i, x_k)\}$, of ordered pairs constructed from a subset of such elements in $X \times X$, denoting the Cartesian product of X with itself. Thus, $(x_i, x_j) \in (X, \succ^*)$ is another way of stating that x_i is chosen over x_j, where $x_i, x_j \in X$. Substituting $\succ^*$ for $\succ$ gives an equivalent way of thinking and denoting arbitrary *preference* patterns (Takemura & Takemura, 2014).

Decision problems and models

In empirical decision theory as it relates to transitivity, a rough distinction can be made between (1) types of decisions problems, (2) decision-making models, and (3) probabilistic models for testing preference axioms (in this case the transitivity axiom).

So far, we have encountered the concept of a choice set X, representing a collection of alternatives that a DM can choose from. In addition, we have used binary relation $\succ^*$ to compactly represent patterns of pair-wise choice over such collections. Implicit in this representation is a certain determinism-of-outcome in relation to a particular choice. That is, for a DM to choose alternative x_i over x_j in a comparison $\{x_i, x_j\}$ is, for all intents and purposes, equivalent to DM receiving some outcome θ_i, where θ_i is identical to x_i. More precisely, the decision problem can be described as a one-to-one mapping, $f\colon X \rightarrow \Theta$, from the set X of alternatives to the set Θ of outcomes. A decision context or problem of this nature is aptly called decision-making under *certainty*, where "certainty" refers to the *epistemic* certainty between an agents action and outcome (Takemura & Takemura, 2014). Within the choice-blindness paradigm, making a decision amounts to a selection between two choice objects, and as far as any DM is concerned, this will result in a determinate outcome (e.g. feedback of the face you selected, political statement you selected, the jam jar you chose). In other words, *what you see is what you get.*

This makes the present study quite unique in that the great majority of experiments investigating transitivity make use of decision-making under *risk* (Birnbaum & Schmidt, 2008; Butler & Pogrebna, 2018; Tversky, 1969).[2] A canonical example of decision-making under risk is betting on a soccer match. For example, betting on a team x_i over x_j is, with some *probability,* associated with an outcome θ_i (you winning some $ amount of money), where the outcome θ_i is dependent on intermediary states of the world (i.e. whether team x_i wins, loses or ties with team x_j). In these contexts, choice of an alternative no longer

[1] Another implication of the 2AFC paradigm is completeness, requiring of a decision maker that she always possesses a well-defined preference for any pair of choice alternatives (i.e., $x_i \succ x_j$, $x_i \prec x_j$, or $x_i \sim x_j$).

[2] Although for interesting exceptions, see Wang, He and He (2021) and Müller-Trede et al. (2015)

has a single one-to-one mapping with an outcome, but is instead only so connected with some *degree* of certainty, where the degree of certainty depends on some probability distribution over worldly states (in this case, the probability of team x_i winning, losing or scoring a tie with team x_j). Likewise in experiments on transitivity, the by far most common stimuli used is different types of gambles or lotteries (Birnbaum & Schmidt, 2008; Cavagnaro & Davis-Stober, 2014; Kalenscher, Tobler, Huijbers, Daselaar & Pennartz, 2010).

The conjunction of a decision context and decision model is enough to predict either transitive or intransitive choice behavior. Decision making under risk has been modelled extensively, the most classical example being that of *Expected Utility Theory* (EUT) (Von Neumann & Morgenstern, 1947). EUT exemplifies a model which not only predicts transitivity but normatively enforces it. It postulates a so called *utility* function U, which gives a real-valued quantitative measure of subjective *value* associated with some set of choice objects (Peterson, 2017). On this account, a preference relation $\succ$ over a choice set X, can always be represented as a *maximization* of utility, such that for any $x_k, x_g \in X$:

(U-representation) $x_k \succ x_g \Leftrightarrow U(x_k) > U(x_g)$

In other words, the preference relation always favors the option with greatest utility (Steele & Stefánsson, 2020). This construct can then by turned into a decision-rule, stating that the best choice (under risk) is always the one which maximizes *expected* utility (EU), where the expected utility for any alternative x_k can be represented by the sum:

$$EU(x_k) = \sum_{i=1}^{n} p_i \cdot U(\theta_i)$$

where p_i is the probability and $U(\theta_i)$ the utility of the i^{th} outcome associated with x_k. Now, since the utility function is defined on the real number line, it must reflect the structure of the greater than (i.e. >) relation on that set, and so transitivity is therefore implied (Briggs, 2019). The numerical representation of preference orderings (as captured by **U-representation**) can be very useful and *will* to some extent be used in the context of both methodology and data analysis in this paper. However, as a descriptive model of choice behavior, EUT has been challenged several times (for a meta-analysis, see Yaqub, Saz & Hussain, 2009); perhaps the most famous case being that of the framing effect as presented by Tversky and Kahneman (1986, 1992). Other transitive decision-models include that of *Cumulative Prospect Theory* (CPT), *Rank Affective Multiplicative* (RAM) and *Transfer of Attention change* (TAX) (Birnbaum, Patton & Lott, 1999; Birnbaum & Schmidt, 2008; Wakker & Tversky, 1993).

A model of risky decision-making which predicts *in*transitive behavior is the *Most Probable Winner* model (MPW). Consider a triple of choice alternatives: x = [15$, 15$, 3$], y = [10$, 10$, 10$], z = [27$, 5$, 5$], each representing a gamble with three equally likely outcomes (i.e. 1/3). According to MPW, a DM will always select a gamble with the highest probability of yielding a favorable outcome (Blavatskyy, 2006). Following this rule, DM selects x on comparison $\{x, y\}$ (x yields favorable outcome 2/3 of the time), y on comparison $\{y, z\}$ (y yields favorable outcome 2/3 of the time), and z on $\{x, z\}$ (z yielding a favorable outcome 5/9 of the time). This implies an *in*transitive choice pattern: $x \succ^* y \succ^* z \succ^* x$. Using 11 sets of this kind of triple, Butler and Pogrebna (2018) has demonstrated that individuals exemplify substantial proportions of intransitive choice patterns.

Risky choice typically involves multi-attribute alternatives that vary in factors such as payoff and probability of success, similar to the gambles mentioned above. But not all multi-attribute choices are risky. Imagine you have rank-ordered preferences for cars along the dimensions of price and horsepower, where horsepower is what you value the most, *all else being equal*. Imagine further that things are *only* equal so long as the price between two cars does not exceed a certain threshold δ. If the difference in price between cars x_1 and x_{1+n} exceed δ, you prefer the cheaper one no matter the difference in horsepower. Given that you follow such a sequential decision-rule, it is possible to construct a sequence of car-comparisons where, some where along the line, you will make a choice that violates transitivity.

In a seminal paper by Tversky (1969) he demonstrates systematic violations of transitive choice by utilizing this kind of structure, technically called *Lexicographic semi-order* (LS). The LS-model is an example of a *non-compensatory* decision strategy, allowing for no trade-off between the attribute values of a choice alternative. The example above demonstrates this lack of trade-off between, in this case, horsepower and price. As soon as one car exceeds the price threshold, information about horsepower is completely disregarded.[3] Other so called *heuristic* decision models, such as Take-the-Best and the *priority heuristic*, are subsets of lexicographic decision strategies (Brandstätter et al., 2006; Gigerenzer & Goldstein, 1996). EUT exemplifies a *compensatory* decision model, describing decision processes where all relevant values are considered and weighted to form a unitary numerical value, which therefore necessitates transitivity.

In a recent study, which takes inspiration from Tversky's original study, Wang et al. (2021) investigated transitivity in the context of human mate preferences. To test how people integrate different cues in partner selection, they constructed partner profiles varying in physical attractiveness (indicated by face pictures) and financial resources. The profile comparisons were set to have a negative correlation between physical attractiveness and financial resources. In theory, this could prompt DMs to apply an intransitivity-compatible non-compensatory cut-off rule. The current study utilizes a similar decision problem with one important exception, namely the (explicit) uni-dimensionality of the choice alternatives, varying only in their degree of physical attractiveness. However, judgements of physical attractiveness is known to decompose (perceptually speaking) into differently prioritised facial characteristics and so the use of lexicographic rules, especially under time constraints, is an ecologically valid possibility (Little, Jones & DeBruine, 2011).

Interestingly however, Wang et al. (2021) found that choice data from the great majority of DMs could best be explained with different stochastic specifications of transitivity. They reference other recent studies which give similar converging evidence for human mate preferences being transitive (Brandner, Brase & Huxman, 2020; Hatz, Park, McCarty, McCarthy & Davis-Stober, 2020). This research gives some clues as to the level of intransitive behavior that will be observed in the

[3] Tversky (1969) conduced two different experiments: one used typical risky gambles varying in probability of success and payoff, whereas the second experiment asked participants choose between hypothetical university applicants, varying in dimensions of intelligence, emotional stability and social facility.

current study, at least in conditions where choice blindness manipulation is not a factor. While the decision problems in this study are not explicitly framed as mate choice problems, it is important to note that assessments of physical attractiveness play a fundamental role in such contexts.

Probabilistic models of Transitivity

A big challenge in the transitivity literature consists in spelling out when and how intransitive preference can be inferred from intransitive choice. An answer to this question is part of what determines if sample data can be said to *truly* violate transitive preference. Contrary to *revealed preference theory* (Samuelson, 1948), choices made on binary comparisons may not always reflect "true" latent preferences. The act of choosing is subject to error and other kinds of variability, a fact which has to be accounted for when studying the empirical validity of a deterministic axiom such as transitivity. There are generally two different approaches to this challenge in the literature. One tries to incorporate variability at the level of binary choice proportions, the other studies aggregate choice *patterns* defined over triples of comparisons (Regenwetter et al., 2010). The former approach is briefly discussed below and subsequently contrasted with the latter.

Tversky (1969) formulated and found violations of a stochastic translation of the transitivity axiom at the level of binary choice proportions, called *Weak Stocastic Transitivity* (WST). Let $p(\psi \succ^* \phi)$ denote the probability of choosing ψ in a $\{\psi, \phi\}$ comparison. WST states that for any (distinct) $x_i, x_j, x_k \in X$,

> **(WST)** if $p(x_i \succ^* x_j) \geq 0.5$ and $p(x_j \succ^* x_k) \geq 0.5$, then $p(x_i \succ^* x_k) \geq 0.5$.

This translation is supposed to account for choice variability and thus allow for inference to *true* intransitive preference if violated. But despite its intuitiveness, WST has well known conceptual issues: repeated rounds of pair-wise choice over a single choice set X, can violate WST if there is a certain mixture of different but all *transitive* choice patterns on that set (Birnbaum & Diecidue, 2015).

In light of this, Regenwetter et al. (2011) suggest an alternative account they call the *mixture model* of transitive preference (MMTP). This model explicitly incorporates variability by allowing mental preferences states themselves to vary from one point in time to another. Nevertheless, choices made at any given time is always the expression of some determinate *transitive* preference order. More precisely, any DM has some (not necessarily uniform) probability distribution over transitive mental states and x_g is chosen over x_k at time t if and only if DM is in a sampled preference state where x_g is preferred to x_k. One such transitive preference state might be $x_i \succ x_g \succ x_k$. This model implies the *Triangle Inequality*, which serves as a supposedly transitive restriction on binary choice proportions akin to that of WST:

> **(TI)** $p(x_i \succ^* x_j) + p(x_j \succ^* x_k) - p(x_i \succ^* x_k) \leq 1$.

This probabilistic translation is immune to the kind of *aggregation* paradox referenced above, i.e. no aggregation of only transitive choice patterns can violate TI. In a reanalysis of Tversky (1969) data (and their own replication of it), Regenwetter et al. (2011) found no significant violation of TI (see also, Cavagnaro & Davis-Stober, 2014).

Unfortunately, TI suffers from the reverse aggregation paradox: an aggregation of *in*transitive choice patterns (and other mixtures) can satisfy TI. That is, in cases where transitive preference should be rejected, TI can indicate the opposite. The moral drawn by Birnbaum from all of this is that "(...) we should be analyzing data patterns rather than marginal choice proportions." (Birnbaum, 2011, p. 5) [4]

The essential idea behind the alternative *pattern counting* approach is to analyse choice data at the triple-level. More precisely, let $\mathcal{T} = \{(T_1, \succ_1^*), (T_2, \succ_2^*), \ldots, (T_i, \succ_i^*)\}$ represent a set of choice patterns for some DM, where each T_i consists of three distinct alternatives. Next, let $\mathcal{T}^{\mathbf{c}}$ be the subset of choice patterns in $\mathcal{T}$ which violate transitivity. For example, $(T_j, \succ_j^*) = \{(x, y), (y, z), (z, x)\}$ would be an element in this set. By counting up the number of elements in $\mathcal{T}^{\mathbf{c}}$ and dividing it by the number of elements in $\mathcal{T}$, one gets to estimate a rate of intransitivity for a given DM. One can then aggregate this rate across participants, which allows for comparison of intransitivity rates between experimental groups or between experimental conditions. This is the main method of analysis that will be used in this study (for other studies using the pattern counting approach, see for example, Butler & Pogrebna, 2018; Loomes, Starmer & Sugden, 1991; Schwartz, Epinat-Duclos, Léone, Poisson & Prado, 2018; Sopher & Gigliotti, 1993).

There are of course caveats with this approach as well. Regenwetter et al. (2011) points out a surprising statistical result regarding the non-monotonic relationship between rates (or *degrees*) of intransitivity and the goodness-of-fit of conventional significance tests. In other words: higher rates of intransitive choice does not necessitate lower p-values. Still, the pattern counting approach avoids blatant conceptual issues befalling both TI and WST. In the present study, comparing rates of intransitivity between experimental conditions gives a straightforward way of revealing the potential effects of choice blindness manipulation on intransitive choice and preference. However, it is in part the aforementioned methodological concerns which motivate novel indirect methods of analysis to be detailed in later sections.

Hypotheses

With the exception for the first one, the following is a fairly abstract level delineation of hypotheses. Only after a detailed account of the methodology can these hypotheses be turned into precise and testable conjectures.

Recall that the main research question (see **(Q)**) intends at a discernment of potential mechanisms for defense against intransitive or otherwise preservation of transitive choice behavior (and by extension transitive preference). Assuming a negative answer to **(Q)** gives *prima facie* reasons for a positive answer to **(q)**, i.e. that choice blindness manipulation can in fact elicit intransitive preference. Under this assumption the first hypothesis can be formulated as follows:

> **(H1)** In conditions of choice blindness manipulation, rates of intransitive preference will be significantly higher when compared to non-manipulated conditions.

In other words: undetected false feedback (and confabulation) will have a marginal effect on the rate of intransitive preference so as to exceed the base rate occurrence in non-manipulated

[4] See also the *True and Error* model, by (Birnbaum & Schmidt, 2008).

conditions. This marginal effect is then to be attributed to *choice blindness induced* (CB-induced) preference change.

Answering **(Q)** in the affirmative results in a set of hypotheses which are both antithetical to the first. One of these hypotheses pertains to defense against choice blindness initiated intransitive choice:

> **(H2a)** Detection of false feedback is more likely when accepting such feedback *implies* an intransitive pattern of choice.

To get a sense for what this means, imagine a DM exemplifying the following choice pattern: $(T_i, \succ^*) = x \succ^* y \succ^* z$. DM therefore chooses x in the forced choice between x and z. False feedback subjected to this choice involves presentation of z as the previously chosen alternative. If DM fails to detect this mismatch and proceeds to give confabulated reasons for choosing z, she has by implication accepted z as the previously chosen alternative, i.e. $z \succ^* x$. This would in turn commit DM to the following *in*transitive choice pattern: $x \succ^* y \succ^* z \succ^* x$. In this sense, detection rate of manipulation is predicted to be sensitive to the *already* defined choice pattern on T_i. Whether the acceptance of false feedback implies intransitivity depends both on the choice pattern exhibited by the DM, and on the specific comparison that is subjected to the manipulation. The details of this fact will be spelled out later.

The second hypothesis builds on the former, but puts emphasis on how a DM might *preserve* transitive preference:

> **(H2b)** An organic *or* CB-induced between-round choice reversal which *would* lead to *in*transitivity, will prompt additional choice reversals to *preserve* a transitive order.

Put simply, a DM will adapt choice behavior accordingly whenever a transitive order on the relevant choice set is threatened. Recall the case described in connection to **H2a**: with $x \succ^* y \succ^* z$ being the initial choice pattern, accepting false feedback on comparison $\{x, z\}$ commits a DM to $z \succ^* x$ and so by implication to the intransitive pattern $x \succ^* y \succ^* z \succ x$. Imagine a case where DM actually reverses her choice on the second repetition of $\{x, z\}$, either due to CB-induced preference change (see *The Choice Blindness Phenomena*) or just by random chance (the organic case). In such cases, a transitive order on T_i is maintained only if DM *also* reverses choice on either the $\{x, y\}$ or $\{y, z\}$ comparisons (or both). For example, an addition reversal on $\{x, y\}$ leads DM to choose y over x, z over x (the reversal which threatened transitivity) and y over z (consistent with original choice in round one). This amounts to the following transitive order on T_i: $y \succ^* z \succ^* x$. As with **H2a**, a more detailed and testable formulation of **H2b** will be formulated in due course.

Finally, it should be noted that **H2a** and **H2b** are not mutually exclusive predictions.

2 Methods

Logical constraints and additional notation

In an effort to induce intransitive preference with false feedback, certain logical imitations are important to keep in mind. Firstly, alternatives of any pair-wise comparison are both elements in a common triple $T_i = \{x, y, z\}$. In an set with three elements there are $\binom{3}{2} = 3$ unique pair-wise comparisons. Therefore, any particular T will be associated with some *order of comparison*, $C = (\{x, y\}, \{y, z\}, \{x, z\})$, representing the pair-wise order in which the elements of T will be shown at the trial-level. As will become apparent soon, deciding on the comparison order for triples is crucial for the manipulation to even be logically compatible with electing intransitivity. We use '$C_i[n]$' to denote the n^{th} term of any C_i.[5] In this experiment, manipulated trials will always coincide with the third and last term of any C_i. If T happens to be a manipulated triple with the order of comparison C, this would mean that subsequent to choice DM is given false feedback on $C[3] = \{x, z\}$. Now, acceptance of false feedback on $C[3]$ only implies intransitivity under circumstances where a DM exemplifies a particular pattern of choice on T, a so called *Choice Blindness Transitive* (CB-Transitive) choice pattern. Informally, a CB-Transitive pattern is one where, *had* the choice made on the last comparison been reversed, the resulting pattern *would* end up *in*transitive. This concept is important in the coming refinement of hypotheses, experimental design and in data analysis.

To get a better grip on CB-Transitivity we can make use of the visual representation of choice patterns in figure 2. The left-most graph (ignoring the dashed arrow) represents choice pattern $(T, \succ_1^*) = x \succ^* y \succ^* z$, black arrows pointing away from the chosen and towards the non-chosen alternative. Assuming C as the comparison order, choice blindness manipulation on $C[3] = \{x, z\}$ can lead to intransitivity if the faulty feedback is accepted (represented by the dashed arrow). The right-most graph represents pattern $(T, \succ_2^*) = x \succ^* z \succ^* y$ (again ignoring the dashed arrow for the moment). In this case, manipulation on $C[3] = \{x, z\}$ *cannot* lead to intransitivity; the counterfactual reversal on $\{x, z\}$ (as represented by the dashed arrow) would result in the following transitive order: $z \succ^* x \succ^* y$. However, reversal of choice on the $\{x, y\}$ comparison *does* lead to intransitivity: $x \succ^* z \succ^* y \succ^* x$. Thus, choice patterns differ crucially on which choice reversal would make them intransitive. We will refer to comparisons in which a reversal of choice results in intransitivity as *target comparisons*.[6] Hence, the target comparison for pattern $(T, \succ_1^*)$ is $\{x, z\}$, whereas for pattern $(T, \succ_2^*)$ the target comparison is $\{x, y\}$.

With this exposition behind us, we can now provide a precise definition of CB-Transitivity. We say that for any choice pattern $(T_i, \succ_i^*)$ and order of comparison C_i,

> **(CB-Transitivity)** $(T_i, \succ_i^*)$ is CB-Transitive if and only if the third comparison $C_i[3]$ is identical to the *target comparison* in C_i, as defined by $(T_i, \succ_i^*)$.

Participants

Participants (N = 118) were recruited using Prolific and were paid 8£ an hour for participation. Inclusion criteria included language proficiency in English and a $>$.95 completion rate. Consent forms were given before and after the experiment. In addition, people that had previously participated in a choice blindness experiment were screened out from the sample.

[5] The subscript 'i' is used both to indicate that we are talking about an arbitrary as opposed to a particular comparison order, but also to associate it to the i^{th} triple T_i.

[6] From a technical standpoint, the elements of a target comparison correspond to the greatest and least elements of T, as determined by the $\succ^*$ relation. An element $\psi \in T_i$ is said to be *the greatest* element of T_i iff for any $y \in T_i$: $\psi \succ^* y$. Correspondingly, an element ϕ is said to be *the least* element of T_i iff for any $z \in T_i$: $\phi \prec^* z$.

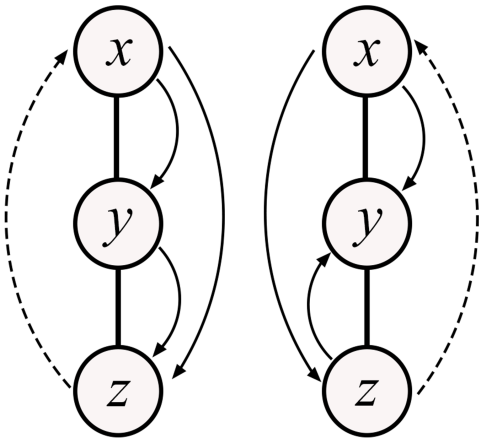

Figure 2: Black and dashed arrows are pointing away from the chosen and towards the non-chosen alternative. The left-most graph is CB-Transitive since choice reversal on the last $C[3]$ comparison (represented by the dotted arrow) would lead to intransitivity. The right-most graph is correspondingly non-CB-Transitive: reversal of choice on $C[3] = \{x, z\}$ leads to the transitive pattern $z \succ^* x \succ^* y$. In this case, reversing the arrow between alternatives x and y would make the pattern intransitive.

Following a pilot study with 15 participants, a frequentist power analysis was conducted with respect to the required sample size to test hypothesis **H1**. To get a plausible effect size estimate, parameters such as detection rates, CB-Transitivity rates and preference change rates were used to calculate an expected difference in the mean intransitivity rate between manipulated and non-manipulated conditions. With the most conservative parameter estimates, this difference amounted to 5%. Hence, the rates of intransitive choice is at worst, expected to be 5% higher in manipulated conditions. Using the sample standard deviation from the pilot data ($\sigma = \sqrt{p(1-p)} = .26$), we end up with a low Cohen's D effect size equal to 0.18. The R-package 'simr' was used to conduct the power analysis for a generalized linear mixed model – to be specified in the section on *Statistical methods* (Green & MacLeod, 2016). With a significance of $\alpha = 0.05$ and a minimum power = 0.80, the minimum sample size to test **H1** is N = 110 (power = 85%, CI[75.26, 92.00]).

Materials

The experiment was implemented online with primary use of plugins from the JavaScript jsPsych library.

Procedure and General Structure

To test stated hypotheses, a within-subject experiment with three experimental phases was conducted. During phase one of the experiment, a DM was asked to sequentially rate pictures of faces using a *visual analogue scale* (VAS). Each face picture was shown for 4.2 seconds and was then covered by a card-back, after which the DM could express her perceived attractiveness of the face. The motivation for using a 0-100 VAS-scale was based on its precision (similar to a slider), its unbiased starting position (requires clicking to initiate tick placement) and the fact that it generates normally distributed data (Funke, 2016). The precision was especially important for current purposes, as the individual ratings of faces were used to determine participant-specific orders of comparison for any choice set (see *Order of comparison*). 20 different triples of face pictures was used, for a total of $20 \times 3 = 60$ trials.

Phase two of the experiment constitutes the main set of binary forced choice trials. On each trial, a DM was tasked with deciding on which of two horizontally aligned faces she finds the most attractive (see figure 1**A**). Both faces pictures are shown for 4 seconds and then covered by differently colored card-backs. DMs then chose the face she preferred by pressing either a 'left face' or a 'right face' button with her mouse cursor. The relative positioning of faces were randomized on each trial, avoiding possible "left/right" face selection bias.

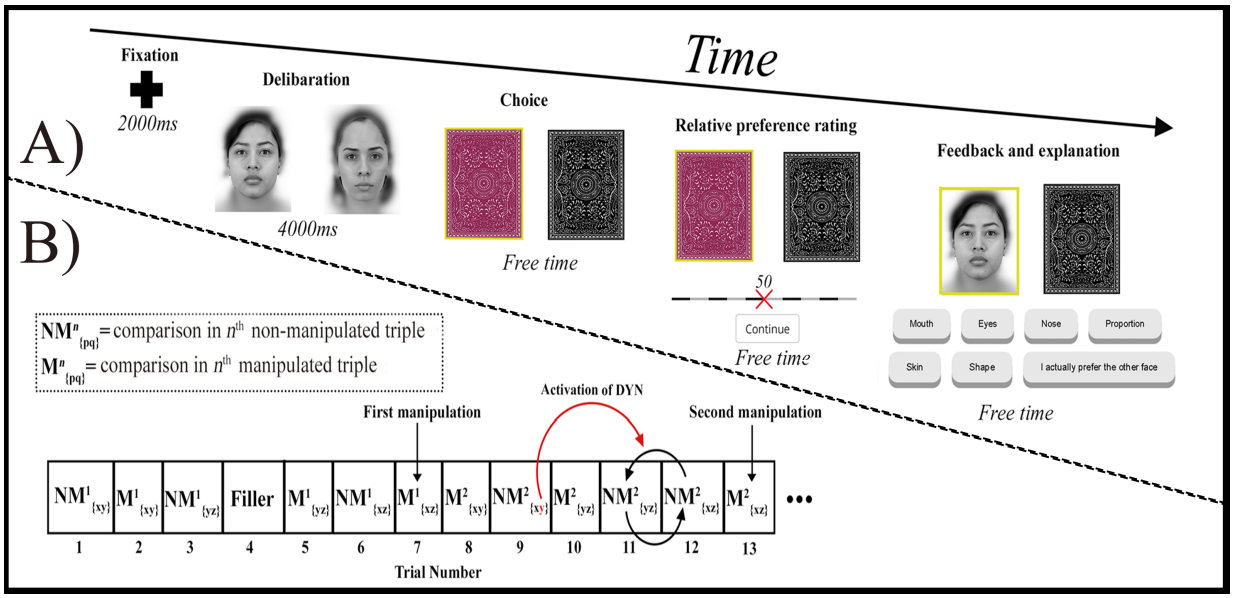

Figure 1: **A** illustrates the main binary choice procedure. In this case, DM receives feedback congruent with her selection. In a manipulated trial DM would be given feedback of the face opposite of her choosing during the **feedback and explanation** screen. **B** represents an example of what the trial-structure may look like. On trial 9, DM selects y and so two pairs from a non-manipulated triple swap places at the trial-level, demonstrating the dynamic presentation rule.

Following this choice, DM stated on a VAS-scale how much stronger her preference was for the face she selected (relative preference rating) (see 1**A**). Followed immediately by this rating was a feedback display showing the *purportedly* selected face picture (in the same spatial position as the 'selected' face). At this stage the face picture was visible until the DM had given an explanation of her choice from a fixed array of facial characteristics (e.g. nose, eyes, proportion). In addition to these options, a DM could also state 'I actually prefer the other face', allowing her to reject false feedback on manipulated trials (or to change her mind) (see figure 1**A**).

In 12 out of 66 trials (including 6 filler trials), DMs were subject to false feedback on the third comparison (i.e. $C[3]$) with respect to a given choice set and its associated comparison order. To reduce detection, the first manipulated trial started at trial position 7-8 and manipulations were then distributed over the remaining trials with a 3-6 trial interval (see figure 1**B**). Fillers were used to ensure this interval is respected. Filler pairs were never manipulated, and so there was likewise 12 cases of manipulations counted at the aggregate triple-level. For each DM, 12 manipulated triples (M-triples) was randomly sampled from the pool of 20, the remaining 8 acted as non-manipulated control triples (NM-triples).

During phase three, participants were tasked with a second round of binary forced-choice trials on the same set of 60 comparisons (filler pairs are excluded). Only this time DMs were not asked to explain their choice subsequent to selection. During this choice phase, the trial structure seen in figure 1**B** was completely randomized, meaning no order was respected, neither internal to a choice set nor between choice sets. By utilizing data from this phase it becomes possible to examine the consistency of choices between the two rounds of binary choice. This data further enables investigation into the marginal effects of manipulation on the rates of intransitive preferences in manipulated conditions (see **H1**).

Stimuli

The stimuli used consisted of hand-picked triples from the Chicago Face Database, a database with standardized photographs and extensive norming data for each individual (e.g. attractiveness ratings, age, sex, race). Attractiveness ratings and algorithmically approximated similarity ratings was used to hand-pick triples of face-pairs. Faces within a triple were broadly similar in their facial features (and by the categories listed above). This effort was mainly made to reduce the detection of false feedback and so boost the number of independent variable manipulations, i.e. *un*detected false feedback on binary choice.

Order of comparison

The order in which triples of comparisons was presented during the first round of binary choice was a crucial component of the experimental design. This was in order to maximize the chances of false feedback being logically compatible with inducing intransitive preferences. Stated with the now introduced vocabulary: we would like to raise the probability that a given choice pattern, $(T, \succ^*)$, satisfies CB-Transitivity relative to its order of comparison C and this amounts to having the target comparison in C coincide with $C[3]$. Now, if one always knew beforehand, exactly what choices a DM was going to make over T, this would be a trivial matter. The crux consists precisely in not having such omniscient powers.

Nonetheless, participant-specific VAS-ratings during the first part of the experiment gives *expected* choice/preference orderings. We can use these ratings to decide which choice object to include in which comparison. More specifically, let f be an *attractiveness rating* function for a DM on a given triple $T = \{x, y, z\}$, such that $f(z) = n$, $f(y) = n + 1$ and $f(x) = n + 2$. The resulting order of presentation for DM will then be given by STAT = $(\{x, y\}, \{y, z\}, \{x, z\})$, i.e. the 1^{st} and 2^{nd} highest rated face is shown first, 2^{nd} and 3^{rd} shown second and the 1^{st} and 3^{rd} shown last.[7] We might call this a *static* rule of comparison since it is based solely on the attractiveness ratings in the first part of the experiment.

If a DMs behavior aligns completely with their prior attractiveness ratings, it would yield a pattern of $x \succ^* y \succ^* z$, since $f(x) > f(y) > f(z)$. This pattern is CB-Transitive since the target comparison $\{x, z\}$ is identical to STAT$[3] = \{x, z\}$.[8] However, if a DM deviates from even one of our prior expectations, this condition will no longer hold. For instance, if DM chooses y over x, the target comparison transforms into $\{y, z\}$, which aligns with STAT[2]. Consequently, any manipulation on STAT[3] could no longer result in intransitivity. The CB-Transitive property is therefore quite fragile.

To increase the likelihood of CB-Transitivity, we used a *dynamic* rule of comparison which was sensitive to the choice made by a DM on the very first comparison in STAT. Let ψ denote the face *selected* on STAT$[1]$ and ϕ denote the face *not* selected on STAT$[1]$. The rule can then be expressed as follows:

$$\text{DYN} = (\{x, y\}, \{\phi, z\}, \{\psi, z\})$$

To understand the advantage of this rule, let's assume, as before, that DM violates f with respect to comparison $\{x, y\}$, choosing y over x instead of the other way around. Despite the prediction error in this case, rule DYN will ensure that DYN[3] coincides with the relevant target comparison: DM selects y in comparison $\{x, y\}$, and so is presented with comparison $\{x = \phi, z\}$ second and $\{y = \psi, z\}$ third. This effectively "saves" CB-Transitivity through rearrangement of term two and three in STAT. Manipulation on the third comparison (if accepted) can now lead to the following intransitive choice pattern: $y \succ^* x \succ^* z \succ^* y$. Figure 1**B** demonstrates the dynamic presentation rule between trials 9-12.

Now, certain prediction errors cannot be accommodated this way, but exactly the same can be said for rule STAT. However, rule DYN will distribute expectancy violations in a more favorable way (with respect to f) across all *possible* transitive choice orders. More precisely, if we assume that it is always more likely to choose in accordance with f than not, then rule DYN will effectively increase the probability, conditional on f, that the third comparison ends up being identical with the target comparison. This is equivalent to saying that DYN raises the probability of any DM satisfying, on any given triple of comparisons, a CB-transitive choice pattern (see Appendix for details).

[7] From now on, variable x, y and z always denotes the highest, second highest and lowest rated face, respectively. Hence, even when a DM exemplifies $z \succ^* y \succ^* x$, the rating function f will always give $f(x) > f(y) > f(z)$.

[8] Recall that manipulations only occur, if and when they occur, on the third and last comparison with respect to any triple. With the introduced notation, we say that manipulations occur *only* on $C_i[3]$ comparisons.

Statistical methods

All statistical modelling and visualization of data was done in R 4.2.0. The primary analysis to test **H1** used a Bayesian mixed logistic regression model, with the proportion of intransitive second-round choice as dependent and choice blindness manipulation as independent variable; subjects and triples as random effects. Every other analysis relating to **H2a** and **H2b** likewise assumed a Bayesian framework, the details of which will be explicated in the results section. Estimates are accompanied with a *Bayes factor* and 95% credible intervals. The Bayes factor (BF) represent the marginal likelihood ratio between the null and the alternative hypothesis (i.e. that there is an effect of some independent variable). For example, BF = 5 means that the sample data is 5 times more likely to be seen under the alternative hypothesis.

3 Results

Refinement of Hypotheses

Having introduced the methodology and additional conceptual machinery, initial formulations of hypotheses can now be refined and made testable. Starting with **H1**, recall that under the assumption of there being no defensive mechanism against intransitivity, the preference change effect is expected to induce some number of intransitive preferences. Thus,

> **(H1*)** In conditions of choice blindness manipulation, rates of second-round intransitive choice will be significantly higher when compared to non-manipulated conditions.

Proceeding to the hypotheses based on the assumption of an affirmative answer to research question **(Q)**, suggesting the presence of a defensive mechanism against intransitivity in cognition. With the concept of CB-Transitivity, hypothesis **H2a** can be formulated as follows:

> **(H2a*)** The detection rate of manipulation will be greater whenever the manipulated comparison belongs to a CB-Transitive choice pattern.

If there is a difference in detection rates between CB-Transitive and non-CB-Transitive choice patterns, it would indicate a higher sensitivity to false feedback when accepting such feedback leads to an intransitive choice.

Before hypothesis **H2b** is reformulated, a preamble is in order. Assume the following trial-level order of comparison for some T_i: $C_i = (\{x, y\}, \{y, z\}, \{x, z\})$. The left-most graph in Figure 3 represents a CB-Transitive choice pattern in *round one* of binary choice trials: $x \succ^* y \succ^* z$. If, in the second round, the choice on $C_i[3]$ is reversed in isolation, intransitivity would follow. Nevertheless, the right-most graph in 3 illustrates (with dotted arrows) how this can be avoided if additional reversals take place on either the $C[1] = \{x, y\}$ or $C[2] = \{y, z\}$ comparisons. Such *conditional* reversals would manifest as drops in choice consistency between the two rounds of binary choice. With this in mind, hypothesis **H2b** can be refined as follows:

> **(H2b*)** Choice consistency on $C_i[1]$ and $C_i[2]$ comparisons will be *uniquely* low when two conditions are met: (1) the round one choice pattern is CB-Transitive and (2) choice is reversed/inconsistent on $C_i[3]$ between the two rounds of binary choice.

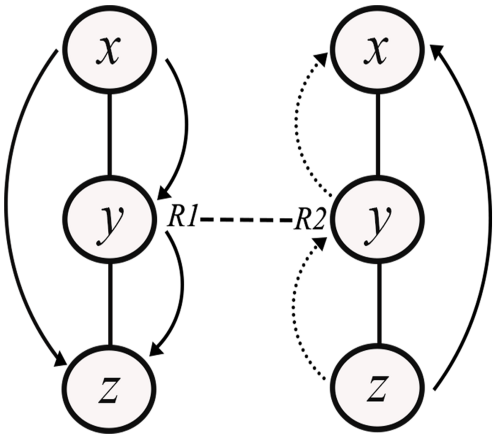

Figure 3: The left-most graph represents the first round (R1) CB-Transitive choice pattern. The right-most graph tries to illustrate, with dotted arrows, the ways for DM to remain transitive in the second (R2), *given* that she has reversed her choice on the $\{x, z\}$ target comparison.

The initial reversal on $C_i[3]$ may be caused by CB-induced preference change (in manipulated conditions) or by organic choice variability. No matter the cause, **H2b*** predicts *systematic* adaptation of non-target choice behavior in order to *preserve* transitivity.

In presenting the results we will use code notation to express different choice/preference patterns. Let numbers 1, 2, and 3 denote choice objects x, y and z, respectively in comparisons $\{x, y\}$, $\{y, z\}$ and $\{x, z\}$. For example, '223' means a DM chose y in $\{x, y\}$, y in $\{y, z\}$, and z in $\{x, z\}$. With two possible choices on each comparison over a total of three different comparisons per triple, there are $2^3 = 8$ possible choice patterns: 121, 131, 133, 123, 221, 233, 223, 231. Patterns 123 and 231 are intransitive, patterns 121 and 221 are CB-Transitive, the rest are non-CB-Transitive.[9]

Rates of Intransitivity

Prior to testing **H1***, it is important to revisit the underlying mechanism that is supposed to drive the elicitation of intransitivity, namely the preference change effect as attributed to choice blindness confabulation (Luo & Yu, 2017; Taya et al., 2014). This effect should manifest as a drop in choice consistency for manipulated pairs, relative to non-manipulated pairs. Figure 4 shows the binary choice consistency between the two rounds of choice, grouped according to conditions of manipulation, detected manipulation, undetected manipulation and non-manipulation. As can be seen by comparing the two middle columns in Figure 4, the total difference in choice consistency between manipulated and non-manipulated conditions is negligible (82.1% vs 82.2%).[10] When consistency data is

[9] Unfortunately, current methodology reduces the transparency of this code, in that the code and the actual trial-level comparison order will not *always* be sequentially aligned. Whenever y is chosen in the first $\{x, y\}$ comparison, $\{x, z\}$ will follow second and $\{y, z\}$ third (in accordance with DYN). So for example, in pattern 221, the first "2" represents selection of y in the first comparison, the second "2" represents selection of y in the third comparison, and "1" represents selection of x in the second comparison.

[10] To ensure the meaningfulness of the comparison between manipulated and non-manipulated conditions, only the two-fold choices made on face-pairs within $C_i[3]$-comparisons are included in both conditions. The rationale behind this is that, on manipulated trials, the manipulation is consistently applied to the $C_i[3]$ comparison and so including other comparisons might lead to con-

aggregated based on CB-Transitivity, it becomes evident that choices belonging to CB-Transitive patterns are more likely to be consistent. However, there is still no meaningful difference observed between the experimental conditions. The consistency rate for CB-Transitive choices is 90% and 91% in M and NM-conditions, respectively. For non-CB-Transitive choices, the consistency rate is 73% in M-conditions and 72.4% in NM-conditions. A comprehensive exploration of the potential reasons for the absence of a preference change effect will be presented in the forthcoming **Discussion**.

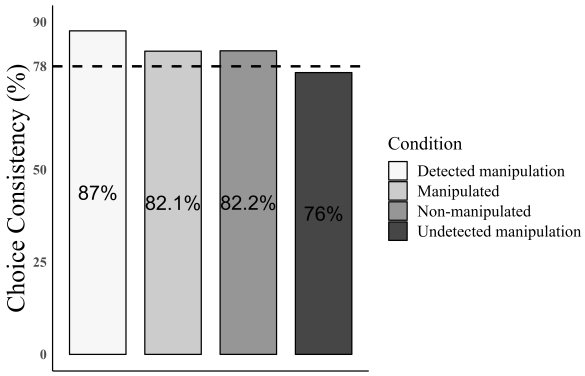

Figure 4: Binary choice consistency between the two rounds of choice, grouped according to conditions of manipulation, detected manipulation, undetected manipulation and non-manipulation. Dotted line represents sample mean across all groupings of the data.

Given these numbers we do not foresee any difference in the rates of intransitive choice between conditions. As expected, according to the logistic regression model specified in *Statistical methods*, 3.5% [2, 5.2] and 3.6% [2.3, 5.23] of choice patterns were intransitive in non-manipulated and manipulated conditions, respectively. The Bayes factor for the (log-odds) β (0.04 [-0.52, 0.58]) was estimated to 0.3. The intransitivity rate was remarkably low in general: averaged across conditions, the intransitivity rate is 3.58% [2.4, 4.8]. Table 1 summarizes the sample proportion of patterns across rounds of choice (R1 vs R2) and condition (NM vs M). For any choice triple T, there are $2^{\binom{3}{2}} = 8$ possible choice patterns, two of which are intransitive (i.e. 123 and 231), and so the intransitivity rate is way below chance-level at 25%.

Table 1: Observed proportion of each pattern, grouped according to round and condition.

	NM-R1	M-R1	NM-R2	M-R2
121	0.33	0.33	0.32	0.34
221	0.18	0.198	0.184	0.166
131	0.18	0.19	0.2	0.19
133	0.08	0.09	0.106	0.092
233	0.06	0.056	0.061	0.066
223	0.08	0.087	0.08	0.084
123	0.029	0.02	0.021	0.029
231	0.037	0.022	0.022	0.025

To get a better idea of the true rate of intransitive *preference*, we follow previous studies in examining the occurrence

founded conclusions about the relative choice consistency.

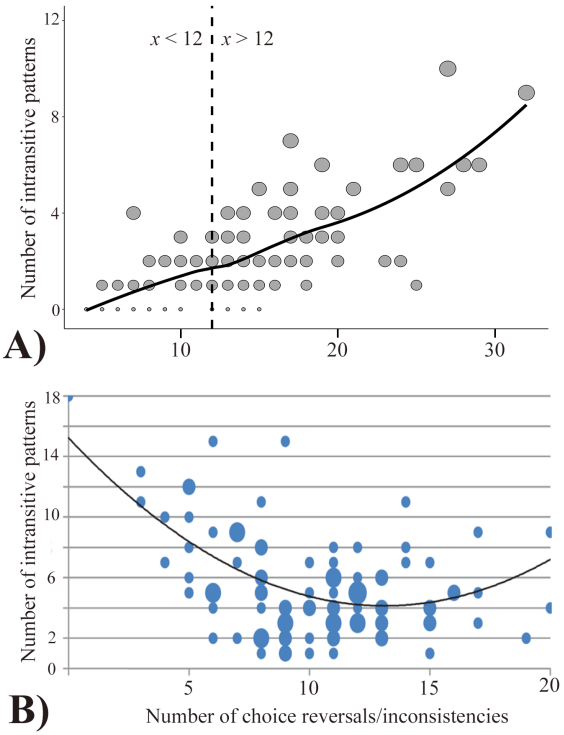

Figure 5: **A** plots individual choice reversals against number of *in*transitive choice patterns. The dotted line divides participants into two even groups based on a threshold of 12 choice reversals. The same analysis from Butler and Pogrebna (2018) can be seen in figure **B**, but with the complete opposite pattern being displayed. *Note.* figure **B** is adapted from "Predictably intransitive preferences", by Butler, D. J., and Pogrebna, G., 2018, *Judgment and Decision Making*, *13*(3), p. 226.

of *repeated* intransitive choice patterns between the two rounds of choice (Birnbaum & Diecidue, 2015; Butler & Pogrebna, 2018). Repeating a choice pattern means consistently selecting the exact same set of three alternatives within a triple across both repetitions of pair-wise choice. When examining all cases of repeated patterns among the total sample size of N = 118, only 2 instances were found to be intransitive, accounting for a mere 1.6%. This contrasts with the findings of the study by Butler and Pogrebna (2018), discussed in the introductory section, where they reported a much higher rate of 32% intransitive patterns among all repeated cases.

To further separate noisy transitivity from true intransitive preference, we follow the aforementioned authors in examining the correlation between individual choice inconsistency (or "noisiness") and the number of intransitive patterns exhibited.[11] In a data set with low rates of repeated intransitivity, we would likewise expect the low rates of intransitive choice (3.58%) to be driven by choice variability rather than true intransitive preference. Figure 5**A** plots the number of choice inconsistencies, at the individual level, against the number of intransitive choice patterns. The dotted line divides the participants into two even groups based on a threshold of 12 choice reversals. The group with <12 reversals exhibited an average of .94 intransitive patterns, whereas participants with >12 reversals exhibited an average of 3.27. The complete reverse pat-

[11] It is important to recall that the choice consistency is identical across the experimental conditions, thus eliminating manipulation condition as a potential confounding factor in the analysis.

tern can be seen in Figure 5**B**, which is an adapted figure from Butler and Pogrebna (2018). In their study, with high rates of intransitive choice and repeated intransitivity, more consistent individuals likewise exhibited greater numbers of intransitive choice patterns.

Detection rate and CB-Transitivity

To get an idea of the general detection rate, a (random slopes) regression model was fit: manipulation condition as fixed effect (two-level factor, NM-conditions coded as intercept), unique DMs and unique face-pair comparisons as random effects. The manipulation-β was estimated to log-odds 4.13 [3.6, 4.67] (BF = $1e^{19}$), which amounts to a detection rate of 46% [36.2, 56.4], meaning that false feedback is accepted roughly 54% of the time. DMs showed large variance in detection rates, with an average standard deviation from the population-β at log-odds 2.22 [1.8, 2.72] (see Figure 6). No such variation was seen with respect to unique face-pairs (log-odds 0.72 [0.41, 1.08]).

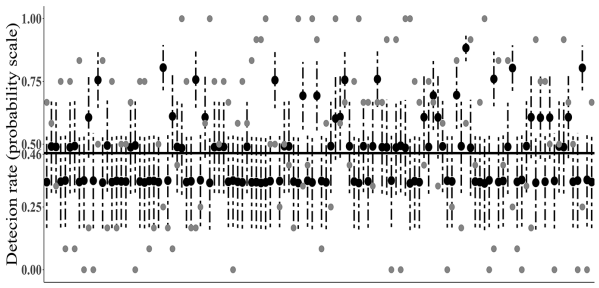

Figure 6: Random effects of unique DMs (N = 118) plotted against individual sample means (grey dots). Dotted line represents model best fit.

Testing **H2a*** involves the control for several confounds. The aim is to study the *isolated* effect of CB-Transitivity on detection rates. Or in other words: the isolated effects of having (hypothetical) *acceptance of false feedback lead to an intransitive choice*. As a first approximation, the sample data showed that manipulation on CB-Transitive patterns resulted in a 20% higher detection rate compared to non-CB-Transitive patterns (CB-Transitive = 60.8%, non-CB-Transitive = 40%). However, this result is only interesting so long as it cannot be wholly explained by preference strength effects.

Firstly, different choice patterns imply different degrees of f-violations, meaning they require of a DM to make different amounts of binary choices that are inconsistent with their previously stated absolute attractiveness ratings. For example, assuming again that $f(1 = x) > f(2 = y) > f(3 = z)$, the CB-Transitive pattern 121 will imply zero violations, whereas the non-CB-Transitive pattern 133 implies two violations, one of which occurs on the last $C_i[3]$ comparison.[12] Sample data indicated that binary choice consistency with f on a manipulated $C_i[3]$ comparison raised the probability of detection by 21% (note that this number may be inflated by other factors). Therefore, difference in preference strength alone might account for the difference in detection rates between CB-Transitive and non-CB-Transitive patterns.

A promising way forward is to compare particular patterns with matching preference strength profiles. Such a comparison can be found between CB-Transitive pattern 221 and non-CB-Transitive 131. 221 and 131 have importantly similar properties except for CB-Transitivity: they have one f-violation each, but are both consistent with f on $C_i[3]$, and yet 221 is CB-Transitive but 131 is not. Comparison of these two patterns amounted to only a 12.6% difference in detection rates (131 = 47%, 221 = 59.6%). However, in this case, there will still be differences in preference strength, but in favor of the non-CB-Transitive pattern. The absolute difference in attractiveness ratings between faces on $C_i[3]$ comparisons is bound to be larger for pattern 131. For pattern 131, $C_i[3] = \{x, z\}$ and for pattern 221, $C_i[3] = \{y, z\}$,[13] and so the mean absolute difference in f-ratings will be greater with respect to 131: $|f(x_i) - f(z_i)| > |f(y_i) - f(z_i)|$. This follows directly from the experimental design of the triple-internal comparison order; faces variables x, y and z always denoting the highest, second highest and lowest rated faces, respectively. Sample data indicated an average difference of 20 rating points for pattern 131, and only 10 points for 221. Similarly, relative preference ratings might differ between patterns. Another measure of preference strength is the consistency of the patterns themselves, i.e. the extent to which they are repeated between the two round of choice. All of these factors are important to control for in forthcoming regression analysis.

To test **H2a***, a mixed multiple regression model was fit on a subset of the data including only manipulated trials belonging to either a 221 or 131 choice pattern. Fixed effects include pattern 221 (131 coded as intercept), difference in f-ratings between faces, relative preference ratings and pattern consistency. Continuous variables are standardized to improve model convergence and interpretation of coefficients (one unit increase equals +1 standard deviation from the mean). Every β-coefficient was allowed to vary with unique DMs, and unique

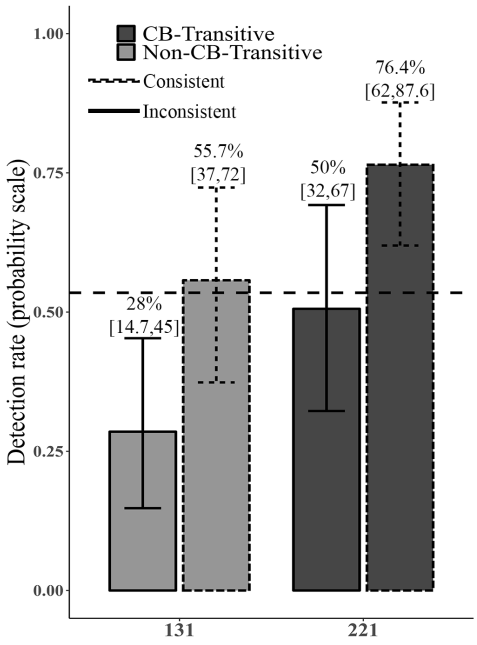

Figure 7: Fitted values from model, grouped according to CB-Transitivity and pattern consistency. Error bars show 95% Bayesian credible intervals. Dotted line represents sample mean across the two patterns.

[12] This is a case where DM chooses z over y and z over x even though $f(x) > f(z)$ and $f(y) > f(z)$.

[13] Recall the dynamic rule of presentation. See the subsection titled *Order of comparison*.

face-pairs were given random intercepts. The number of 121 and 221 patterns were evenly balanced (280 cases of 121 and 268 cases of 131).

Results indicated a strong and significant effect of CB-Transitivity on detection rates (log-odds 0.98 [0.39, 1.57], BF = 48.61). Computation of the fitted posterior contrast between pattern 221 and 131 (averaged across pattern consistency) showed that DMs were 21.4% [8, 34] more likely to detect false feedback whenever the manipulated comparison belonged to a 221 pattern (see Figure 7). A moderate effect of confidence ratings can be seen (log-odds 0.53 [0.15, 0.97], BF = 9.7), where $+1\sigma$ from the mean predicts roughly 11.7% higher detection rates. With respect to the effect of attractiveness difference the evidence is in favor of the null (log-odds 0.09 [-0.23, 0.42], BF = 0.2). Pattern consistency was also a strong predictor of higher detection rates, but is more uncertain than the CB-Transitivity-β (log-odds 1.19 [-0.23, 2.10], BF = 13.7).

In sum, a substantial part of the variance cannot be accounted for by reference to differences in preference strength. Thus, what drives higher detection rates for pattern 221 is its CB-Transitive property. In other words: DMs are markedly more sensitive to false feedback when accepting such feedback entails an intransitive pattern of choice.

Evidence for Transitivity preserving preference change

Recall that for CB-Transitive choice patterns, isolated choice reversals on $C_i[3]$ comparisons inevitably results in intransitivity. According to **H2b***, such cases will prompt additional choice reversals on $C_i[1]$ *or* $C_i[2]$ in order for a DM to *preserve* a transitive order on the associated triple. In terms of modelling, we are therefore interested in the interaction between CB-Transitivity and $C_i[3]$ consistency as it is related to choice consistency on $C_i[1]$ and $C_i[2]$ comparisons. When a pattern is non-CB-Transitive, $C_i[3]$ inconsistency should have no considerable effect on the $C_i[1]$ and $C_i[2]$ choice consistency. In these cases, reversal on $C_i[3]$ does *not* threaten a transitive order on the associated choice set. As in the case with **H2a***, the aforementioned interaction effect would only be interesting so long as differences in preference strength cannot account for substantial parts of the variance.

The bottom panel in Figure 8 shows relative preference ratings for $C_i[1]$ and $C_i[2]$ comparisons, grouped according to CB-Transitivity and $C_i[3]$ consistency. Across all combinations of CB-Transitivity and $C_i[3]$ consistency, no pronounced difference can be seen (Non-CB-Transitive-consistent: mean = 46.2; CB-Transitive-consistent: mean = 47.6; Non-CB-Transitive-inconsistent: mean = 49.3; CB-Transitive-inconsistent: mean = 46.9). The top panel in Figure 8 represents absolute differences in attractiveness ratings between faces in $C_i[1]$ and $C_i[2]$ comparisons, grouped in the same way as before. In this case, there was a slight advantage for comparisons belonging to CB-Transitive choice patterns (Non-CB-Transitive-consistent: mean = 10.5; CB-Transitive-consistent: mean = 12.9; Non-CB-Transitive-inconsistent: mean = 11; CB-Transitive-inconsistent: mean = 12.6).

As mentioned previously, making a binary choice consistent with attractiveness ratings is another good measure of preference strength and was shown to be predictive of higher detection rates. Likewise in the case of choice consistency, choosing in accordance with f-ratings was predictive of 12.8% higher between-round choice consistency (f-consistent = 81.6%, f-inconsistent = 69%). In other words: a DM who

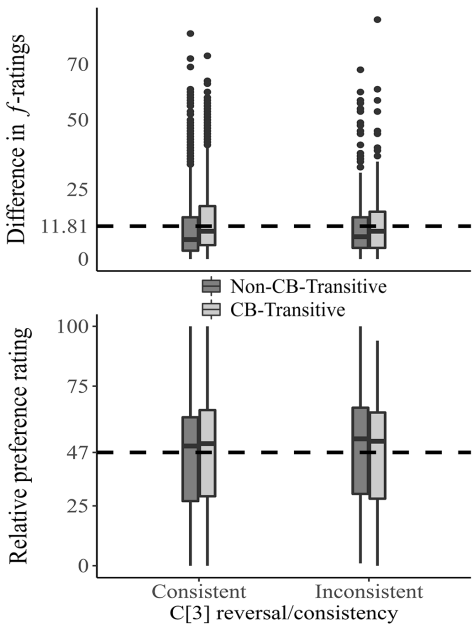

Figure 8: Plots the relative preference ratings (bottom) and absolute difference in f-ratings (top) as a function of $C_i[3]$-consistency and CB-Transitivity. The dotted line in both panels represents the respective sample means.

chooses a face with higher attractiveness rating, according to her own previous judgement, will be more likely make the same selection this face again. As also previously mentioned: the experimental design necessitates that CB-Transitive patterns are more consistent with f-ratings compared to non-CB-Transitive patterns (see Appendix for details), implying that CB-Transitive choices are driven by stronger subjective preferences. Consequently, if choice consistency is driven entirely by preference strength and is completely insensitive to the abstract property of transitivity, higher consistency rates for comparisons in CB-Transitive patterns should be expected across the board. In fact, in the section titled *Rates of intransitivity*, it was shown that CB-Transitive choices (aggregated across all comparisons) had higher choice consistency rates in both M and NM-conditions.

To test **H2b*** a mixed interaction model was fit on a subset of the data including only $C_i[1]$ and $C_i[2]$ comparisons. To repeat, the outcome variable of interest is binary choice consistency on the just referenced comparisons. Fixed effects include CB-Transitivity, $C_i[3]$-consistency and CB-Transitivity*$C_i[3]$-consistency as an interaction term. All coefficients were allowed to vary with unique DMs. The model estimated a moderate positive effect of CB-Transitivity (log-odds 0.28 [0.10, 0.46], BF = 8.42). However, as expected under **H2b***, this effect was reversed when co-present with a reversal on a $C_i[3]$ comparison, as shown with a decisive interaction-β estimated to log-odds -1.58 [-2.04, -1.11]; BF = $1.9e^5$.

Figure 9 plots the fitted values from the model. When choice was consistent on $C_i[3]$ comparisons, $C_i[1]$ and $C_i[2]$ consistency was higher for CB-Transitive patterns (i.e. either 121 or 221). The fitted posterior contrast was 4.3% [1.6, 7.6]. In line with the interaction-β, the complete opposite effect of CB-Transitivity becomes apparent when there is a reversal on $C_i[3]$. The fitted posterior contrast indicated a drop in choice consistency by 26% [17.2, 34.9]. In the non-CB-Transitive

case, $C_i[3]$-inconsistency is, if anything, associated with an uncertain increase in choice consistency (posterior contrast: 5% [0.4, 9.7]; $C[3]$-consistency-β = log odds 0.32 [0.02, 0.65], BF = 1.46). In sum, strong evidence has been given in favor of **H2b***: despite the presence of a strong initial subjective preference, cross-round choice dynamics exhibits systematic adaptation in line with the abstract property of transitivity.

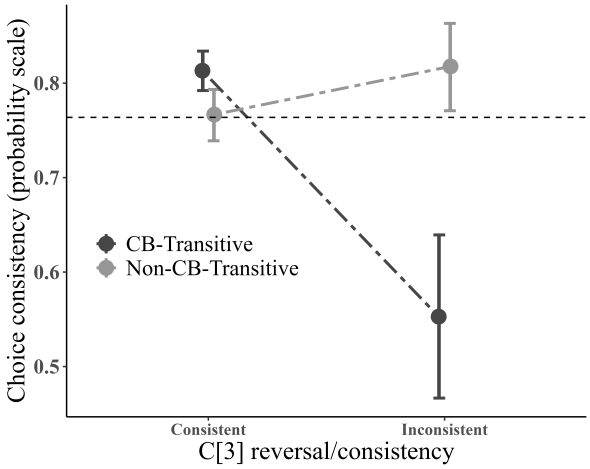

Figure 9: Fitted values from interaction model. This figure has the same structure as figure 8. Labels on the x-axis indicate whether choice is consistent (between rounds) on the $C_i[3]$ comparison. The y-axis represents choice consistency for $C_i[1]$ and $C_i[2]$ comparisons. Dotted line represents sample mean.

4 Discussion

Elicitation of intransitivity — Limitations

At its face, the results indicate it *not* being possible to induce intransitivity with choice blindness. However, equally apparent is the absence of CB-induced *preference change*, an otherwise robust effect replicated in multiple studies (Izuma et al., 2015; Johansson et al., 2014; Luo & Yu, 2017; Taya et al., 2014). For instance, in Johansson et al. (2014) the difference in choice consistency between M (56.6%) and NM-conditions (93.3%) was roughly 37%, whereas in this study the difference approaches zero. One interesting possibility is that the effect of preference change is somehow resisted to prevent the emergence of intransitive preferences. More specifically, when common CB-Transitive patterns are subject to manipulation, the otherwise seen preference change effect might be "blocked" by some transitivity preserving mechanism. However, in that case we should still expect see the normal effect of preference change in non-CB-Transitive cases. That is, when preference change is no threat to transitivity, there ought to be a normal discrepancy between M and NM-conditions. No such indications can be found in the data. Since preference change is the driving mechanism behind the hypothesized effect, **H1** cannot be properly rejected with current data.

Several factors may account for the failure to see a preference change effect, some more interesting than others. Firstly, diminished engagement and self-reflective interaction with the chosen face is believed to be on of the contributing factors. The "explanation" of each face-selection consists in a simple button press from an array of facial characteristics. Although the "selected" face is visible until an explanation has been given, it is possible that participants pay more attention to the buttons themselves and are disposed by convenience to make a heedless response. Taya et al. (2014) found an expected preference change effect in an online-implemented study where participants had to provide typewritten explanations for their choice (on half of the trials). In a follow-up study, the general procedure will be adapted in a similar way, with the hope that this makes *confabulation* genuine and more effective in aligning later choice with false feedback. This may also improve the baseline choice consistency (78%, see Figure 4), another factor which may dampen the effect size in this case.

Secondly, the current experiment is adapted for the study of transitivity, which introduces certain structural features uncommon to other choice blindness experiments. Recall that the initial triple-internal comparison order (barring the dynamic rule) is based on participant-relative attractiveness ratings. Couple this fact with patterns 121 and 131 being two of the most common patterns in round one M-conditions (see table 1). Both of these patterns contain faces, in the last manipulated $C[3]$ comparison, that has the lowest and the highest attractiveness ratings in their respective triple. Hence, it is possible that the *differentiability* of these face-pairs partly hinders the otherwise observable shift in second round choice, following undetected manipulations. In other words, the "confabulation"-threshold for successfully inducing preference change, is elevated by having a large part of the manipulations befall face-pairs with relatively large discrepancies in attractiveness.

Thirdly, detection rates of manipulation are relatively high (10%-30% detection rates seen in previous work, see for example, Johansson et al., 2005; Taya et al., 2014). It is known from previous studies as well as this one (see Figure 4) that *detected* manipulations increase choice consistency between rounds (e.g. Johansson et al., 2014). Low detection rates is therefore a prerequisite for CB-induced preference change to be discernible between experimental conditions. In a more speculative vein, high detection rate for a given participant (say, above 50%) may have a global negative effect on the minority of undetected cases, in the sense of nullifying the otherwise effective confabulation mechanism. Lowering detection rates going forward may involve making faces more similar (in attractiveness and on other dimensions). At the same time, such efforts can unfavorably lower the baseline rate in choice consistency; states of indifference promotes choice variability. Another alternative is to exclude the very first phase of the experiment. Reason being that previous exposure and explicit rating of faces plausibly raises detection rates on binary choice. However, this approach would significantly hinder ones ability (as a researcher) to construct optimized comparison orders, which is crucially important when trying to maximize the rates of CB-Transitivity. Future investigation should explore whether meta-data (e.g. averaged attractiveness ratings) can act as a viable alternative for constructing triple-internal orders.

Intransitivity in General

Two lines of reasoning allow for confident rejection of intransitive *preference* in the data. In line with other researchers, the rate of *repetition* of a pattern serves as a good approximation to whether it constitutes a *true* preference structure (Birnbaum & Diecidue, 2015; Butler & Pogrebna, 2018). The presumption is that *true* preference patterns exhibits robustness to choice variability or "error", and that repetition of the same pattern twice means you've demonstrated such robustness. In this study,

practically no repeated intransitive patterns are found. This is solid grounds for attributing the few cases of intransitivity to choice variability with underlying transitive preferences.

Under the assumption that intransitive choice is reflective of noisy transitivity, greater magnitudes of noise is expected to, purely by virtue of stochastic processes, generate greater magnitudes of intransitivity. The correlation seen between choice inconsistency and intransitivity rates therefore gives converging evidence to the same conclusion. The analysis represented in figure 5**B** from Butler and Pogrebna (2018), found the exact opposite correlation on a data set with very high rates of intransitive choice and *repeated* intransitive choice. They found that participants with fewer choice reversals (i.e. more consistent individuals) had significantly higher rates of intransitive choice, supporting the rest of their analysis in favor of *truly* intransitive preferences. If their reasoning is valid, the precise opposite conclusion can be drawn from the data in this study, i.e. the few intransitive patterns observed are *not* true cases of intransitive preference.

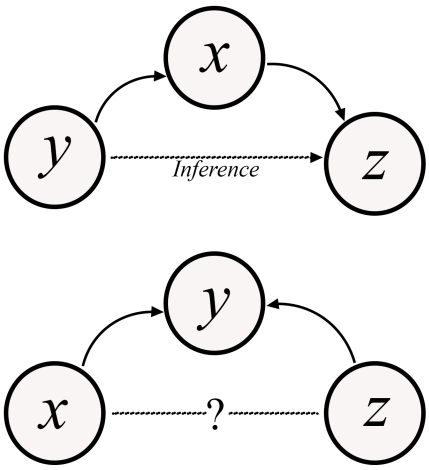

Figure 10: The top graph represents a condition where transitive inference can take place. The graph below allows for no such inference. Black arrows represent actual choices made and are pointing *away* from the chosen alternative.

Transitive Inference

The detection rate analysis strongly corroborates **H2a**: participants are substantially more likely to notice false feedback when the acceptance of such feedback *would* compel them to *in*transitive choice. This is a novel kind of analysis and result in the behavioral decision-theoretic literature. As such, it will require quite speculative efforts in its interpretation.

One avenue looks to the well recognized phenomena of *transitive inference* (TI). This refers to the cognitive capacity of being able to infer $x\mathbf{R}z$ from the prior knowledge or exposure to $x\mathbf{R}y$ and $y\mathbf{R}z$, where $\mathbf{R}$ is a transitive relation on some arbitrary set of objects. In a typical transitive inference (TI) task, subjects (human or non-human animals) make reinforced or trial-and-error based discriminatory judgements on sequences of item-overlapping comparisons: $\{x+, y\}, \{y+, z\}, \{z+, b\}...$ — where "+" represents the in some sense "better", "correct" or otherwise rewarded alternative. The *critical* test is then made on a novel non-adjacent comparison, such as $\{x, z\}$, where successful selection of x is taken as evidence for a transitive inference (Greene, Spellman, Levy, Dusek & Eichenbaum, 2001). TI has been extensively demonstrated empirically in adult, non-adult, and other non-human animals (e.g. monkeys, rodents, fish) (Bryant & Trabasso, 1971; Davis, 1992; Frank, Rudy & O'Reilly, 2003; Heckers, Zalesak, Weiss, Ditman & Titone, 2004; Koscik & Tranel, 2012; McGonigle & Chalmers, 1977). Experimental evidence indicates that the TI-capacity presupposes no explicit logical reasoning or meta-conscious awareness on the part of the agent (Frank, Rudy, Levy & O'Reilly, 2005; Greene et al., 2001). Within the current decision-context, it is plausible that a process akin to implicit transitive inference occurs at times when DMs exemplify CB-Transitive choice patterns.

To see why, we need to further elucidate certain properties of pattern 221, the CB-Transitive pattern compared with 131 in the analysis of detection rates. Pattern 221 denotes the selection of y in the first $\{x, y\}$ comparison. Because of the dynamic rule of presentation (see **Methods**), comparison $\{y, z\}$ and $\{x, z\}$ will swap places at the trial-level: $C = (\{x, y\}, \{x, z\}, \{y, z\})$. A DM then chooses $x = 1$ on the now second comparison $\{x, z\}$ and $y = 2$ on $\{y, z\}$, the third and last comparison. Notice that for a perfectly rational agent, having made the first two choices makes the last $\{y, z\}$ comparison superfluous; there is only one choice which is consistent with transitivity, namely y. The first two choices in pattern 131 come with no such implications: choosing x on $\{x, y\}$ and z on $\{y, z\}$ leaves open a choice of either x or z on the last $\{x, z\}$ comparison (resulting in patterns 131 and 133, respectively). Thus, the logical relationship between the two initial choices in pattern 221 is hypothesized to directly initiate *or* support a future inferential process concerning the "correct" choice in the final $\{y, z\}$ comparison, where the "correct" choice is defined as the one that conforms to a transitive choice order. The two scenarios just described are depicted in Figure 10. The top graph demonstrates the inference thought to be initiated in the case of pattern 221, the bottom graph demonstrates the lack of such an inference in the case of pattern 131. How does this help us understand the discrepancy in detection rates?

An answer is found in the close conceptual and cognitive relationship between inference and *prediction*. In the context of TI, the inferential process is essentially a future-directed prediction with respect to your own future choice behavior, and I claim, with respect to the *consequences* of those actions. The idea is the following: when a DM chooses y in the last comparison (in pattern 221), there is an *expectation* of receiving y as feedback and this expectation is grounded both in having actually chosen y, but *also* in its *transitive* relationship to the other two choices within the same choice structure. Thus, when a DM is given false feedback on the aforementioned choice, not only is there an outcome-mismatch but a prediction-error relating to the expected transitive order on the associated choice set. Although a very rough sketch, thinking along these lines can help make sense of the robust difference in detection rates between pattern 221 and 131.

This kind of inferential reasoning might rely on quite sophisticated relational representations of the elements in a choice set.[14] More specifically, at the time of deliberation and choice it would seem to require memory retrieval of discrete event elements followed by recombination of these elements to sup-

[14]Note that patterns 221 and 131 have structurally identical reinforcement histories. On the crucial last comparison in both patterns, only one alternative (y in 221 and x in 131) has been "reinforced" previously by being selected in the first comparison. Explanation via reinforcement histories can therefore not explain the discrepancies in detection rates between pattern 221 and 131.

port inference about future events and actions. Precisely these kinds of flexible representations are thought to be supported by the hippocampus, hypothesized to be able to rapidly code individual event details and their relationships to one another (O'Reilly & Rudy, 2001). Partial inputs (e.g. display of the last $\{y, z\}$ comparison) are thought to trigger *pattern completion* mechanisms, leading to the retrieval of previous associations between elements (e.g. "y was chosen over x" and "x was chosen over z") that can support inferences about anticipated future associations ("y *will* be chosen over z") (Eichenbaum & Fortin, 2009; Zeithamova, Schlichting & Preston, 2012). Although speculative, these kinds of processes might underlie a kind of *cementation* of the transitively inferred alternative to the rest of the choice structure (in this case face y), in turn prompting stronger anticipations about the future outcomes of action.

In sum, it has been argued that a general capacity for transitive inference and prediction can play a plausible incidental role in defending against intransitivity in choice and preference. This defensive mechanisms, to repeat, manifests as discrepancies in detection rates between CB-Transitive and non-CB-Transitive patterns of choice.

Transitive choice variability

With respect to **H2b**, what the results show is that variability in choice is not random. When participants change their preference or otherwise display variability, they do so in a manner that is consistent with transitivity. This is not a trivial point. Under the assumption that decision-makers are indifferent to transitivity, it becomes difficult to explain why only certain choice reversals should trigger *additional* reversals on comparisons which are otherwise quite consistent. In other words, a plausible account of the interaction seen in Figure 9 involves taking seriously the implications of the surface-level behavior, namely that it systematically preserves transitivity.

The *mixture model* of transitive preference by Regenwetter et al. (2011) offers a possible way to make sense of current data. As briefly mentioned in the introduction, on this account binary choice always originates from transitive preference states, but the particular preference state associated with a choice is allowed to vary from one time to another. Again, let P($\psi \succ^* \phi$) denote the probability of choosing ψ in a comparison $\{\psi, \phi\}$. Then, for a given triple T, let Γ denote the set of all strict linear orders on T for a given DM. The linear orders are strict since each element in Γ is defined from a strict preference relation $\succ$ on T. Mathematically, the mixture model states that P($x \succ^* y$) for some DM is equal to:

$$P(x \succ^* y) = \sum_{\substack{\succ \in \Gamma \\ x \succ y}} P_{\succ}$$

where $P_{\succ}$ is the total probability of a DM being in a preference state $\succ$ where x is preferred to y (i.e. $x \succ y$). In words, the overt probability that a DM selects x in a $\{x, y\}$ comparison is given by the probability that she is in a mental preference state where x is preferred to y. The idea behind this model is that decision-makers draw, at different points in time, preference orders from some probability distribution over transitive mental states.

Now, the psychological plausibility of this idea depends on the interval at which preference states are supposedly sampled: is a new preference state randomly drawn at each new binary decision-problem, or only when the decision-maker is in a sufficiently different context? The former case would seem to entail that each binary choice is made somehow independently of any other, even in cases where choice objects belong to the same choice set. This is not only hard to believe but would contradict the earlier discussion on transitive inference, based entirely on the idea that previously established logical relationships between elements can be grounds for negotiating future decision-making.

Building on the mixture model, Müller-Trede et al. (2015) has adopted precisely the idea of context-dependent preferences. In these so called *context-sensitive preference models*, the probability attached to a preference state can depend on present and past choice contexts. The present context is defined (at least partly) by the currently available choice alternatives and their attributes. In Sher and McKenzie (2014) "options-as-information" model, DMs use currently available choice alternatives and their attributes to create "posterior" models of a wider range of possible choice alternatives and their attributes. This posterior attribute distribution is then part in what determines the probability distribution over preference states, which in turn determines pair-wise choice. According to Müller-Trede et al. (2015), the *triangle inequality* is no longer a normative requirement in context sensitive preference models, thus avoiding previously discussed conceptual issues with the original version of the mixture model (see the section titled *Probabilistic models of Transitivity*).

This experiment provides one *salient* difference in contexts between the two rounds of choice, pertaining specifically to the *order* in which different alternatives are sampled. Recall from the **Methods** section that the comparison order in the second round of choice is completely randomized. This means that comparisons, within a triple, that were previously encountered first, may now be encountered last or second. For example, the $\{x, y\}$ comparison which in round one is always presented first, may now be presented last relative to the other comparisons within the triple. Coupled with constrained memory of previous choices, it therefore becomes plausible that, with differing orders of sampled options, decision-makers likewise sample distinct yet fully *transitive* preference orders on the second round of choice. To investigate this further one could start with two statistical tests: (1) are choice patterns more likely to be repeated when the alternatives within a triple are shown in the same order during both rounds of choice? (2) Do particular comparison orders correlate with particular choice patterns?

Present considerations does not establish a rigorous case for the aforementioned models, but it does give an intuitive interpretation of the systematic choice reversals seen in Figure 9. Going forward, more extensive analysis could be done with respect to individual choice patterns and their cross-round dynamics. For every transitive choice pattern there exists exactly one *target comparison*, which *if* reversed in isolation, *would* lead to intransitivity. Take patterns 131 and 223 for example: an isolated between-round choice reversal on the first $\{x, y\}$ comparison would make both of these patterns intransitive in the second round (from 131 to 231 and from 223 to 123, respectively). In fact, any of the three possible comparisons (i.e. $\{x, y\}$, $\{y, z\}$, $\{x, z\}$) have two *uniquely* associated transitive patterns, such that choice reversal on any of these comparisons would make two associated patterns intransitive. Therefore, we can equally ask what happens to non-target choice consistency in these cases: will similar transitivity preserving choice reversals obtain?

5 Concluding remarks

Experimental findings lend no evidence to the presence of structurally intransitive preferences. This holds both in relation to CB-induced intransitivity and with respect to any natural occurrence of intransitive choice. However, given certain methodological limitations, the possibility of CB-induced intransitivity remain open until further adapted investigation has been conducted.

Evidence is given for different kinds of *transitivity preserving* behavior. Empirical data and educated speculation suggests that a predictive mechanism rooted in transitive inference can account for discrepancies in detection rate between CB-Transitive and non-CB-Transitive choice patterns. In addition, analysis of cross-round choice dynamics reveal systematic efforts to adapt future choice behavior in line with the transitivity axiom. It is suggested that the results can be interpreted within *context sensitive mixture models* of transitive preference.

Acknowledgements

I would like to extend my gratitude to Gabriel Vogel and Petter Johansson for letting me take part in a very exciting project and for giving me invaluable guidance throughout my work. I am looking forward to future collaborations where present methodological limitations can be addressed. I am indebted to Sonja Holmer and Mia Huovilainen for their valuable comments on an earlier draft of this thesis. Finally, I would like to thank Daniel Carlström Schad for fruitful discussions and for lending sources of inspiration in interpreting current experimental results.

References

Auvray, M., Gallace, A., Hartcher-O'Brien, J., Tan, H. Z. & Spence, C. (2008). Tactile and visual distractors induce change blindness for tactile stimuli presented on the fingertips. *Brain research*, *1213*, 111–119.

Bar-Hillel, M. & Margalit, A. (1988). How vicious are cycles of intransitive choice? *Theory and decision*, *24*, 119–145.

Bernoulli, D. (1738). Exposition on a new theory on the measurement of risk. *Econometrica*, *22*(1), 23-36. (Translated by Dr. Louise Sommer, January 1954)

Birnbaum, M. H. (2011, Oct). Testing mixture models of transitive preference: comment on regenwetter, dana, and davis-stober (2011). *Psychological Review*, *118*(4), 675–683. doi: 10.1037/a0023852

Birnbaum, M. H. & Diecidue, E. (2015). Testing a class of models that includes majority rule and regret theories: Transitivity, recycling, and restricted branch independence. *Decision*, *2*(3), 145.

Birnbaum, M. H., Patton, J. N. & Lott, M. K. (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, M. H. & Schmidt, U. (2008). An experimental investigation of violations of transitivity in choice under uncertainty. *Journal of Risk and Uncertainty*, *37*, 77–91.

Blavatskyy, P. R. (2006). Axiomatization of a preference for most probable winner. *Theory and Decision*, *60*, 17–33.

Block, W. E., Barnett, I. et al. (2012). Transitivity and the money pump. *Quarterly Journal of Austrian Economics*, *15*(2).

Brandner, J. L., Brase, G. L. & Huxman, S. A. (2020). "weighting" to find the right person: Compensatory trait integrating versus alternative models to assess mate value. *Evolution and Human Behavior*, *41*(4), 284–292.

Brandstätter, E., Gigerenzer, G. & Hertwig, R. (2006). The priority heuristic: making choices without trade-offs. *Psychological review*, *113*(2), 409.

Briggs, R. A. (2019). Normative Theories of Rational Choice: Expected Utility. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2019 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/fall2019/entries/rationality-normative-utility/`.

Bryant, P. E. & Trabasso, T. (1971). Transitive inferences and memory in young children. *Nature*, *232*, 456–458.

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

Cavagnaro, D. R. & Davis-Stober, C. P. (2014). Transitive in our preferences, but transitive in different ways: an analysis of choice variability. *Decision*, *1*(2), 102.

Davis, H. (1992). Transitive inference in rats (rattus norvegicus). *Journal of Comparative Psychology*, *106*(4), 342.

Dickerson, K. & Gaston, J. R. (2014). Did you hear that? the role of stimulus similarity and uncertainty in auditory change deafness. *Frontiers in psychology*, *5*, 1125.

Eichenbaum, H. & Fortin, N. J. (2009). The neurobiology of memory based predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1521), 1183–1191.

Frank, M. J., Rudy, J. W., Levy, W. B. & O'Reilly, R. C. (2005). When logic fails: Implicit transitive inference in humans. *Memory & Cognition*, *33*(4), 742–750.

Frank, M. J., Rudy, J. W. & O'Reilly, R. C. (2003). Transitivity, flexibility, conjunctive representations, and the hippocampus. ii. a computational analysis. *Hippocampus*, *13*(3), 341–354.

Friedman, M. & Savage, L. J. (1952). The expected-utility hypothesis and the measurability of utility. *Journal of Political Economy*, *60*(6), 463–474.

Funke, F. (2016). A web experiment showing negative effects of slider scales compared to visual analogue scales and radio button scales. *Social Science Computer Review*, *34*(2), 244–254.

Gigerenzer, G. & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological review*, *103*(4), 650.

Green, P. & MacLeod, C. J. (2016). simr: an r package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. Retrieved from `https://CRAN.R-project.org/package=simr` doi: 10.1111/2041-210X.12504

Greene, A. J., Spellman, B. A., Levy, W. B., Dusek, J. A. & Eichenbaum, H. B. (2001). Relational learning with and without awareness: Transitive inference using nonverbal stimuli in humans. *Memory & cognition*, *29*(6), 893–902.

Hatz, L. E., Park, S., McCarty, K. N., McCarthy, D. M. & Davis-Stober, C. P. (2020). Young adults make rational

sexual decisions. *Psychological Science*, *31*(8), 944–956.

Heckers, S., Zalesak, M., Weiss, A. P., Ditman, T. & Titone, D. (2004). Hippocampal activation during transitive inference in humans. *Hippocampus*, *14*(2), 153–162.

Izuma, K., Akula, S., Murayama, K., Wu, D.-A., Iacoboni, M. & Adolphs, R. (2015). A causal role for posterior medial frontal cortex in choice-induced preference change. *Journal of Neuroscience*, *35*(8), 3598–3606.

Johansson, P., Hall, L. & Sikström, S. (2008). From change blindness to choice blindness. *Psychologia*, *51*(2), 142–155.

Johansson, P., Hall, L., Sikstrom, S. & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), 116–119.

Johansson, P., Hall, L., Tärning, B., Sikström, S. & Chater, N. (2014). Choice blindness and preference change: You will like this paper better if you (believe you) chose to read it! *Journal of Behavioral Decision Making*, *27*(3), 281–289.

Kalenscher, T., Tobler, P. N., Huijbers, W., Daselaar, S. M. & Pennartz, C. M. (2010). Neural signatures of intransitive preferences. *Frontiers in Human Neuroscience*, *4*, 49.

Kivetz, R. & Simonson, I. (2000). The effects of incomplete information on consumer choice. *Journal of marketing research*, *37*(4), 427–448.

Koscik, T. R. & Tranel, D. (2012). The human ventromedial prefrontal cortex is critical for transitive inference. *Journal of cognitive neuroscience*, *24*(5), 1191–1204.

Little, A. C., Jones, B. C. & DeBruine, L. M. (2011). Facial attractiveness: evolutionary based research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1571), 1638–1659.

Loomes, G., Starmer, C. & Sugden, R. (1991). Observing violations of transitivity by experimental methods. *Econometrica: Journal of the Econometric Society*, 425–439.

Loomes, G. & Sugden, R. (1987). Some implications of a more general form of regret theory. *Journal of Economic Theory*, *41*(2), 270–287.

Luce, R. D. (1995). Four tensions concerning mathematical modeling in psychology. *Annual Review of Psychology*, *46*(1), 1–27.

Luce, R. D. (1997). Several unresolved conceptual problems of mathematical psychology. *journal of mathematical psychology*, *41*(1), 79–87.

Luo, J. & Yu, R. (2017). The spreading of alternatives: Is it the perceived choice or actual choice that changes our preference? *Journal of Behavioral Decision Making*, *30*(2), 484–491.

McGonigle, B. O. & Chalmers, M. (1977). Are monkeys logical? *Nature*, *267*, 694–696.

Müller-Trede, J., Sher, S. & McKenzie, C. R. (2015). Transitivity in context: A rational analysis of intransitive choice and context-sensitive preference. *Decision*, *2*(4), 280.

Nishimura, H. & Ok, E. A. (2018). *Preference structures* (Tech. Rep.). mimeo, NYU.

O'Reilly, R. C. & Rudy, J. W. (2001). Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychological review*, *108*(2), 311.

Peterson, M. (2017). *An introduction to decision theory*. Cambridge University Press.

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

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

Samuelson, P. A. (1948). Consumption theory in terms of revealed preference. *Economica*, *15*(60), 243–253.

Schwartz, F., Epinat-Duclos, J., Léone, J., Poisson, A. & Prado, J. (2018). Impaired neural processing of transitive relations in children with math learning difficulty. *NeuroImage: Clinical*, *20*, 1255–1265.

Sher, S. & McKenzie, C. R. (2014). Options as information: Rational reversals of evaluation and preference. *Journal of Experimental Psychology: General*, *143*(3), 1127.

Simons, D. J. & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in cognitive sciences*, *9*(1), 16–20.

Sopher, B. & Gigliotti, G. (1993). Intransitive cycles: Rational choice or random error? an answer based on estimation of error rates with experimental data. *Theory and Decision*, *35*, 311–336.

Steele, K. & Stefánsson, H. O. (2020). Decision Theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2020 ed.). Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/win2020/entries/decision-theory/`.

Strandberg, T., Sivén, D., Hall, L., Johansson, P. & Pärnamets, P. (2018). False beliefs and confabulation can lead to lasting changes in political attitudes. *Journal of Experimental Psychology: General*, *147*(9), 1382.

Takemura, K. & Takemura, K. (2014). Behavioral decision theories that explain decision-making processes. *Behavioral Decision Theory: Psychological and Mathematical Descriptions of Human Choice Behavior*, 143–164.

Taya, F., Gupta, S., Farber, I. & Mullette-Gillman, O. A. (2014). Manipulation detection and preference alterations in a choice blindness paradigm. *PloS one*, *9*(9), e108515.

Tversky, A. (1969). Intransitivity of preferences. *Psychological review*, *76*(1), 31.

Tversky, A. & Kahneman, D. (1986). Judgment under uncertainty: Heuristics and biases.

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

Von Neumann, J. & Morgenstern, O. (1947). Theory of games and economic behavior, 2nd rev.

Wakker, P. & Tversky, A. (1993). An axiomatization of cumulative prospect theory. *Journal of risk and uncertainty*, *7*, 147–175.

Wang, H., He, Z. & He, L. (2021). Transitive mate preferences. *Decision*, *8*(3), 180.

Yaqub, M. Z., Saz, G. & Hussain, D. (2009). A meta analysis of the empirical evidence on expected utility theory. *European Journal of Economics, Finance, and Administrative Sciences*, *15*, 117–133.

Zeithamova, D., Schlichting, M. L. & Preston, A. R. (2012). The hippocampus and inferential reasoning: building memories to navigate future decisions. *Frontiers in human neuroscience*, *6*, 70.

Appendix

The tables below show all the possible transitive choice patterns given either a STAT or DYN rule of comparison (6 possibilities each). The leftmost column shows the number of expectancy violationsfor each sequence given $f(x) > f(y) > f(z)$, expressing that x is rated higher than y, y higher than z and x higher than z. In this context it will be helpful to use a choice function g to represent the binary choice made on any given comparison. g takes a comparison $\{x_i, x_j\}$ and returns a subset $\{x_i\}$ representing the chosen alternative. The code notation introduced before is likewise used in the tables below. Again, note that this code will not always be aligned with the comparison order under the dynamic rule, since comparison $\{y, z\}$ and $\{x, z\}$ swap places whenever a DM selects y in comparison $\{x, y\}$. A '$\star$' will be used to indicate whenever the there is swap of comparisons under the dynamic rule.

DYN

f-violations	Comparison 1	Comparison 2	Comparison 3	CB-transitive	Code
0	$\{x,y\} = \{x\}$	$\{y,z\} = \{y\}$	$\{x,z\} = \{x\}$	Yes	121
1	$\{x,y\} = \{y\}$	$\{x,z\} = \{x\}$	$\{y,z\} = \{y\}$	Yes	221 $\star$
1	$\{x,y\} = \{x\}$	$\{y,z\} = \{z\}$	$\{x,z\} = \{x\}$	No	131
2	$\{x,y\} = \{x\}$	$\{y,z\} = \{z\}$	$\{x,z\} = \{z\}$	No	133
3	$\{x,y\} = \{y\}$	$\{x,z\} = \{z\}$	$\{y,z\} = \{z\}$	No	233$\star$
2	$\{x,y\} = \{y\}$	$\{x,z\} = \{z\}$	$\{y,z\} = \{y\}$	No	223$\star$

Table 2: Transitive preference orderings under rule DYN. Function symbol g is left out for notational brevity.

STAT

f-violations	Comparison 1	Comparison 2	Comparison 3	CB-transitive	Code
0	$\{x,y\} = \{x\}$	$\{y,z\} = \{y\}$	$\{x,z\} = \{x\}$	Yes	121
3	$\{x,y\} = \{y\}$	$\{y,z\} = \{z\}$	$\{x,z\} = \{z\}$	Yes	233
1	$\{x,y\} = \{x\}$	$\{y,z\} = \{z\}$	$\{x,z\} = \{x\}$	No	131
2	$\{x,y\} = \{y\}$	$\{y,z\} = \{y\}$	$\{x,z\} = \{z\}$	No	223
1	$\{x,y\} = \{y\}$	$\{y,z\} = \{y\}$	$\{x,z\} = \{x\}$	No	221
2	$\{x,y\} = \{x\}$	$\{y,z\} = \{z\}$	$\{x,z\} = \{z\}$	No	133

Table 3: Transitive preference orderings under rule STAT. Function symbol g is left out for notational brevity.

Relative to rule STAT, we can see that rule DYN makes it less likely, given f, that a DM satisfies any of the ordinary non-CB-Transitive patterns (8 vs 6 total f-violations), and *more* likely that a DM satisfies any of the two CB-transitive patterns (1 vs 3 total f-violations). This is why rule DYN is better than rule STAT; it raises the conditional probability, given f, that S will satisfy a CB-transitive choice pattern.