A Stress-Test of Economic Rationality

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Felix Jan Nitsch

aus Düsseldorf

Düsseldorf, September 2021

aus dem Institut für Experimentelle Psychologie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

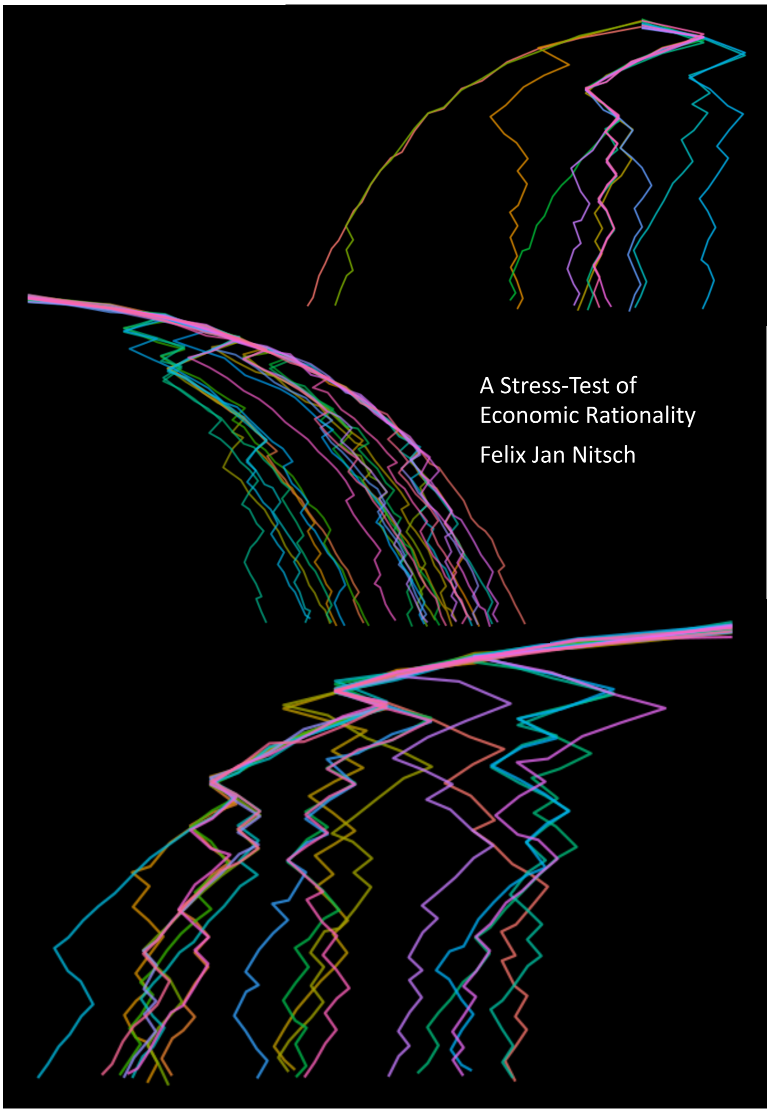
Berichterstatter:

1. Prof. Dr. Tobias Kalenscher

2. Prof. Dr. Gerhard Jocham

Tag der mündlichen Prüfung: 02.03.2022

Cover Art



The cover art of this dissertation represents the Garden of Forking Paths. It tries to capture the path of our lives, created by our choices, that took us to where we are – but also where we could be if we made different choices. I created the cover in R – and there's an accompanying R Shiny App where you can create colorful decision trees yourselves (https://fjnitsch.shinyapps.io/DecisionArt/).

Danksagung

Eine Promotion zum Thema rationaler Entscheidungen ist ein kniffliges Unterfangen, da der typische, unvermeidbar einsetzende Selbstzweifel durch die ständige Beschäftigung damit, was gute Entscheidungen ausmacht, in ungünstiger Weise verstärkt wird. Glücklich für mich war, dass ich während dieser innerlichen Auseinandersetzung auf zahlreiche Resilienz-Faktoren bauen konnte, und ich so überwiegend doch recht entspannt blieb. Entscheidend war sicherlich auch, dass mir die eigentliche Arbeit stets Freude bereitete, auch spätabends oder am Wochenende.

Ich danke meinem Betreuer, Tobias Kalenscher, für diese spannende Herausforderung und erfahrungsreiche Zeit. Die Promotion hat mir sicherlich nicht nur wissenschaftlich das nötige Rüstzeug für meine weitere Laufbahn vermittelt. Weiterhin möchte ich meinem Zweitbetreuer Gerhard Jocham danken, welcher mir gerade zu Beginn und zum Ende der Promotion einige sehr hilfreiche Ratschläge gab.

Ich danke meiner Familie – Mama, Papa, Martin und Georg – für die emotionale, intellektuelle und pragmatische Unterstützung, sowie die Ablenkung an richtiger Stelle. Ich weiß, dass ich mich immer auf euch verlassen kann. Ebenfalls danke ich meiner Verlobten Lorena, welche meine Arbeitswut, meine Selbstgespräche im Home-Office und meine Unfähigkeit, Urlaub zu machen tapfer ertrug, und mich stets daran erinnerte, mich nicht völlig in der Arbeit zu verlieren.

Ich danke meinen Kolleginnen und Kollegen, insbesondere Douman, Manu, Maurice, Sandra, und Yue, sowie meinen Koautorinnen und Koautoren für die produktive Zusammenarbeit, den intellektuellen Austausch und die Gespräche in der Mittagspause. Auch möchte ich den zahlreichen Studierenden, studentischen Hilfskräften und

Versuchsteilnehmenden danken, ohne deren praktische Unterstützung viele Projekte nicht möglich gewesen wären. Ich danke allen meinen Freundinnen und Freunden bei NEURD, insbesondere Sami und Hannah, für den Austausch und das konstruktive Feedback, stets in bester Atmosphäre. An dieser Stelle sei auch Wissenschafts-Twitter gedankt, welches mich mit offener Wissenschaft vertraut machte, aktuelle Forschungsarbeiten rezensierte und mir einen kleinen Einblick in die Kultur der Akademie gewährte.

Weiterhin möchte ich meinen Mentorinnen und Mentoren beim *Rheinischen FührungsColleg*, sowie meiner Stipendiaten-Gruppe danken, welche mir in der Schlussphase der Promotion einen spannenden Außenblick auf meine Arbeit und meinen Werdegang ermöglichten.

Zuletzt danke ich der Mensa und *Huel* für meine Nährstoffversorgung, sowie Death Metal für den nötigen Drive – laut Spotify gehörte ich stellenweise zu den Top-0.5% Hörern der Band *In Flames*, welche insgesamt durchschnittlich ca. 1.7 Millionen monatliche Hörer verbucht. \m/

Hiermit schließe ich diesen Lebensabschnitt und freue mich auf viele neue Erfahrungen, Herausforderungen und spannende Forschungsergebnisse bei meiner nächsten Etappe in Frankreich.

Zusammenfassung

Die Frage, wie man gute oder rationale Entscheidungen trifft, beschäftigt Philosophen, Wissenschaftler und Praktiker seit Jahrhunderten bis zum heutigen Tag. Ökonomische Rationalität im Speziellen kann gemäß der Erwartungsnutzentheorie als die Fähigkeit definiert werden, stets die subjektiv beste Option für sich selbst zu wählen - was mit einer konsistenten Entscheidungsfindung unter Kosten einhergeht. In den Arbeiten, über die in dieser Dissertation berichtet wird, haben wir die Erwartungsnutzentheorie als normativen Maßstab verwendet, um vergleichende, nicht absolute Aussagen über die Rationalität von Entscheidungen zu treffen. Konkret haben wir mehrere Experimente durchgeführt, um drei potenzielle Einflussfaktoren auf Rationalität (unklare Zielvorstellungen, akuter Stress und chronischer Stress) zu identifizieren, sowie eine qualitative und quantitative Analyse der Literatur durchgeführt, um den aktuellen Stand der Forschung zusammenzufassen. Schließlich haben wir ein grundlegendes methodisches Validierungsexperiment zur Messung der Entscheidungskonsistenz durchgeführt, dessen Ergebnisse potenziell weitreichende Konsequenzen für die heutige Forschungspraxis haben. Insgesamt deuten unsere Ergebnisse darauf hin, dass Entscheidungskonsistenz weder ein robustes noch ein verlässliches Merkmal von Entscheidungsträgern ist. Aber unsere empirische Arbeit zeigt auch, dass nicht jede Störung (z. B. akuter Stress) unbedingt zu einer verminderten Rationalität führen muss. Darüber hinaus verdeutlicht unsere Arbeit, dass ökonomische Konzepte aus theoretischen und praktischen Gründen nicht naiv mit psychometrischen Maßen gleichgesetzt werden sollten. Diese Dissertation trägt zu einem aktuellen Forschungsprogramm in der Neuroökonomie bei, welches Faktoren identifiziert, die die Entscheidungsqualität beeinträchtigen könnten.

Schlüsselwörter: Rationalität, Erwartungsnutzentheorie, offenbarte Präferenzen, Stress

Abstract

The question of how to make good or rational decisions has puzzled philosophers, scientists, and practitioners for centuries until today. According to expected utility theory, economic rationality, specifically, can be defined as the capacity to always choose the subjectively best option for oneself – which coincides with choosing consistently under cost. In the work that is reported in this dissertation we used EUT as a normative benchmark to make comparative, not absolute statements about rationality of choice. Specifically, we conducted experiments to identify three potential influence factors on rationality (unclear choice goals, acute stress, and chronic stress), as well as a qualitative and quantitative analysis of the literature body to summarize the current state of research. Lastly, we conducted a foundational methodological validation experiment on choice consistency measurements, whose results have potentially far-reaching consequences for the contemporary research practice. Our findings tentatively suggest that choice consistency is neither a robust nor reliable trait of decision makers, but our empirical work also highlights that not every nuisance (i.e. acute stress) must immediately lead to reduced rationality. Further, our work highlights that economic concepts ought not be naively mistaken for psychometric measures for theoretical and practical reasons. Our work contributes to a research program in neuroeconomics that strives to identify factors that could compromise decision quality.

Keywords: rationality, expected utility theory, revealed preference, stress

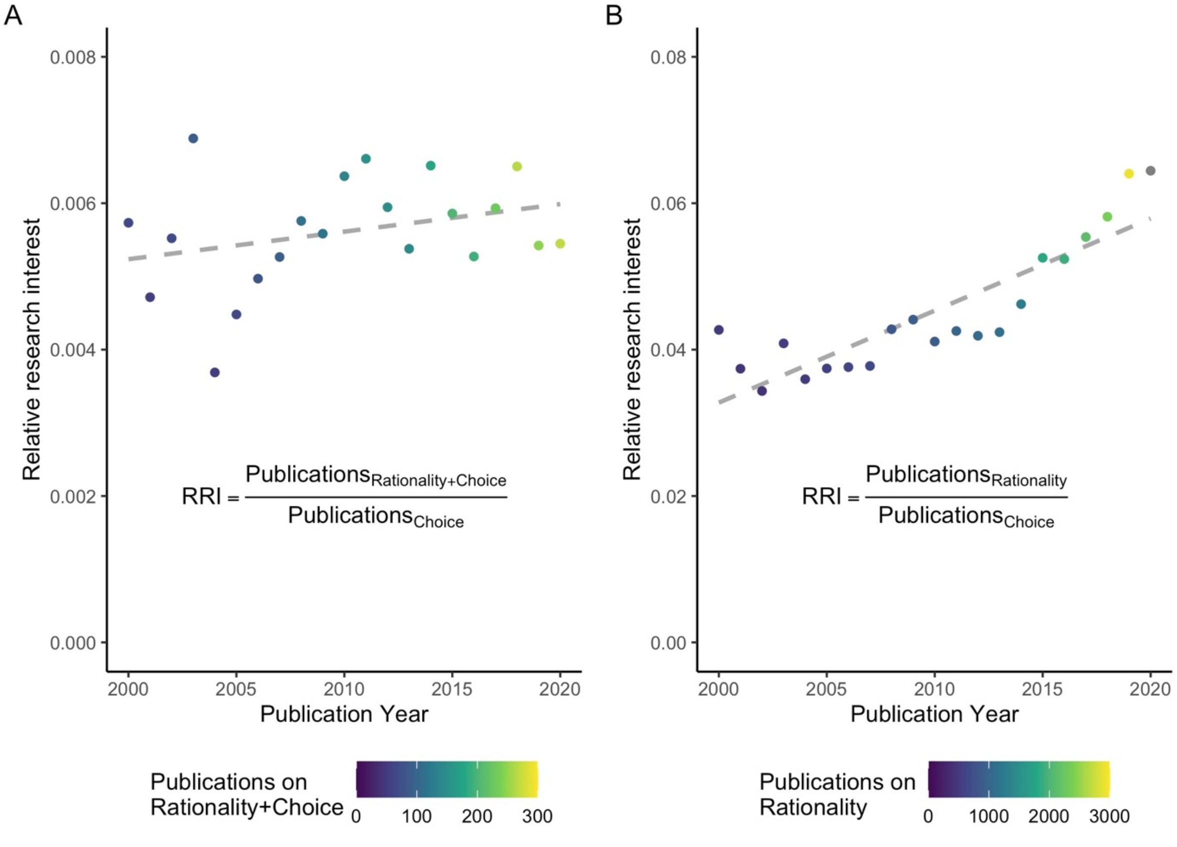
TABLE OF CONTENTS

General Introduction

Whether in education, romantic relationships, or financial matters – we often find ourselves pondering our future decisions. Most of us probably also know the feeling of regret after making a presumably wrong choice. Our decisions have far-reaching consequences for our and others' lives, and the question of how to make good or *rational* decisions has puzzled philosophers, scientists, and practitioners for centuries until today (see figure 1). Hence, it is not so surprising, perhaps, that various differently-nuanced conceptualizations of rationality exist. Understanding these nuances is critical to understanding and evaluating the ongoing discourse in the various fields adjacent to rational choice theory.

Etymologically, the terminus rationality derives from the Latin origin *rationalitas*, which can be translated as the capacity to think or reason (in German: Denkvermögen or Vernunftbegabung). In philosophy, there have been many attempts to define reason by reference to presumably objective and undeniable values, for example by Aristotle or Kant. A reasonable act could then be simply defined as an act in alignment with these values (c.f. Horkheimer, 1946). In modern psychology and economics, however, the prevailing conceptualization of reason and, thus, rationality follows a different, subjective and instrumental tradition. Here, rationality is defined as the alignment and logical application of means to ends and the reasonableness of actions is fully contingent on the subjective preference (Bentham, 1780; Hume, 1738; Mill, 1863). In the perhaps most hyperbolic form of this view of rationality, Hume famously argued that the preference for the "destruction of the world to the scratching of my finger" (Hume, 1738, p. 416) is rational, as would be any other preference per se. Henceforth, when talking about rationality I will solely refer to this latter definition. Of course, this view of rationality is not uncontested (see e.g. Horkheimer, 1946), but the philosophical discourse on this matter is beyond the scope of this dissertation.

Figure 1: Research Interest in Rationality



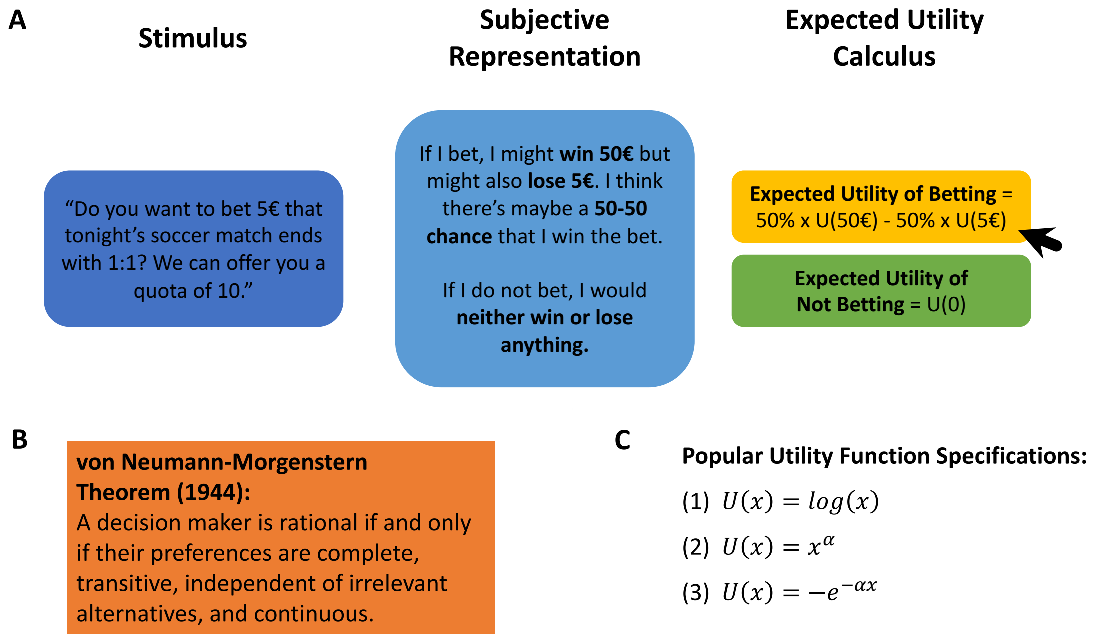
Please note the scale difference between panel A and B

Depicted is the trend of the relative research interest in rationality over the last two decades. Relative research interest was operationalized as the number of publications on the topic (panel A: “rationality”; panel B “rationality AND choice”) divided by the number of publications on choice in general (“choice”) found on Web of Science. This normalization is necessary to control for the trend of increasing publications in all fields (as also apparent in the color coding). Depending on which operationalization is used, relative research interest in rationality remained constant or tentatively increased.

Expected utility theory

The modern mathematical axiomatization of economic rational choice theory was provided in the 1950s in the form of expected utility theory (EUT; see figure 2, panel B; Morgenstern & Von Neumann, 1953). EUT states that rational agents should always choose that one out of a set of alternatives, for which the greatest utility is expected (see figure 2, panel A).

Figure 2: Expected Utility Theory



"Do you want to bet 5€ that tonight's soccer match ends with 1:1? We can offer you a quota of 10."

If I bet, I might **win 50€** but might also **lose 5€**. I think there's maybe a **50-50 chance** that I win the bet.

If I do not bet, I would **neither win or lose anything.**

Expected Utility of Betting = 50% x U(50€) - 50% x U(5€)

Expected Utility of Not Betting = U(0)

von Neumann-Morgenstern Theorem (1944): A decision maker is rational if and only if their preferences are complete, transitive, independent of irrelevant alternatives, and continuous.

Popular Utility Function Specifications:

(1) $U(x) = log(x)$

(2) $U(x) = x^{\alpha}$

(3) $U(x) = -e^{-\alpha x}$

Depicted is expected utility theory (EUT) in a nutshell. Panel A shows the decision-making process: An external stimulus (e.g. a gamble) is subjectively represented as probabilities and outcomes which form the input of the expected utility calculus on the right. The choice option with the highest expected utility is selected. Panel B summarizes the core axioms of EUT according to the von Neumann-Morgenstern theorem. Panel C names a few examples of popularly used utility functions. The form of the utility function has implications for the behavioral response to risky prospects.

The standard interpretation of utility here is a scalar value that is a function of the agent's subjective preferences (see figure 2, panel C). In economics (with the exception of neuroeconomics, which often assumes that utility is represented in the brain) it is generally interpreted as a non-psychological entity (Gul & Pesendorfer, 2008) and, thus, not directly measurable. Importantly, EUT imposes no restrictions on the direction of preferences. Consequently, the *expected* utility of a choice denotes the sum of the utilities of all possible consequences of a choice times their respective probabilities of realization. Since Savage (1972), probabilities in EUT are considered subjective, usually in a Bayesian sense.

A classic metaphor for the type of choices considered in EUT are gambles. For example, an agent might decide between betting on the result of a coin flip, where heads and tails are associated with a monetary consequence depending on the choice, e.g. winning 5€ if the chosen side shows. If we assume a fair coin where each side has a 50% chance to show, then, the expected utility of choosing heads in this thought experiment would be the sum of the utilities of winning and not winning 5€ divided by two.

Revealed preference theory

EUT can also be represented via choice consistency axioms. A brilliant recognition of 20th century economics was that choice consistency is a necessary and sufficient condition for rationality (Afriat, 1973; Houthakker, 1950; Samuelson, 1938; Varian, 1982). This at least partly solved the measurement problem of utility for economists outlined above. Following Varian (2006):

> **Definition 1.** If a choice option is selected over another available option, it is directly revealed preferred (RD).
>
> **Definition 2.** If a choice option is selected over another available option at a cost, it is strictly directly revealed preferred (SD).
>
> **Generalized Axiom of Revealed Preference (GARP).** Let R be the transitive closure of RD. Then, for all x, y where x R y there is no x, y where y SD x.
>
> **Afriat's Theorem (partial).** If a set of choices satisfies GARP, there exists a non-satiated, continuous, monotone, and concave utility function that rationalizes the data.

Note, that with Definition 2 we have, in passing, introduced the concept of cost (i.e. positive prices). In a market situation, the prices of goods are thought to vary according to their supply and demand. To meaningfully speak of choices at a cost, it is necessary that the decision-maker either has an explicit understanding of prices, where each choice option is unambiguously associated with a price, or an implicit understanding of prices, where the price of an option can be inferred from the properties of the choice option. Under such circumstances, Afriat's Theorem implies cost efficiency (Afriat, 1972). Using Afriat's Theorem, we arrive at a common contemporary operationalization of EUT, which is also a cornerstone of this dissertation:

Definition 3. Rationality of choice means to choose consistently under cost.

A classic argument to establish the validity of rational choice theory sensu choice consistency is the *money pump*: in an economic context, "an arbitrageur would be able to extract money from an inconsistent agent indefinitely, without providing any services in return, by presenting him or her with a carefully chosen sequence of trades" (Cubitt & Sugden, 2001, p. 121). For example, if an agent agrees to trade back and forth two goods, each time at a cost, such inconsistency of preference would ultimately deprive them of all their wealth. By transitivity, this argument holds for circular sequences of costly trades of any length. A similar argument (*Dutch books*) can be made for that subjective probabilities in EUT must follow the laws of probability theory (Ramsey, 1926; see Vineberg, 2016 for an accessible overview; see Appendix A).

Interpretations of rational choice theory

EUT can be – and historically has been interpreted in multiple ways: as a descriptive theory, it makes statements about *how and why* people make decisions; as a predictive theory, it makes statements about which decisions people *do* make; as a normative theory, it makes statements about which decisions people *should* make. While all interpretations of EUT have received substantial criticism (e.g. Harless & Camerer, 1994; Rieskamp et al., 2006; Tversky, 1975), especially the normative interpretation of EUT as a benchmark of decision quality still has wide traction in the applied disciplines (Corner & Kirkwood, 1991; Huang et al., 2011; Keefer et al., 2004; Velasquez & Hester, 2013). Generally speaking, it is useful to adopt a perspective of scientific instrumentalism, where we embrace that EUT, like any theory, is not true or false, but rather a means to an end, of which the latter we must define at the start of our research work (see General Discussion). Specifically, in the work that is reported in this dissertation we used EUT as a normative benchmark to make comparative, not absolute statements about rationality of choice.

Background and research questions

This dissertation mostly contributes to a contemporary research program in neuroeconomics that identifies factors, which could potentially compromise decision quality (for an exemplary outline see the introduction of Choi et al., 2014). The research program is neuroeconomic in the sense that it is interested in the influence not only of economic and demographic but also neuropsychological factors (e.g. intelligence, brain structure and activity, memory, sleep, neuroendocrine modulators etc.) on economic rationality as an indicator of decision quality. However, our work is explicitly agnostic on whether expected utility is itself represented neuropsychologically. Specifically, we conducted experiments to identify three novel potential influence factors (unclear choice goals, acute stress, and chronic stress), as well as a qualitative and quantitative analysis of the literature body to summarize the current status of the research agenda. Lastly, we conducted some foundational methodological work on choice consistency measurements, that have potentially far-reaching consequences for contemporary research practice.

The above-mentioned line of research bears relevance beyond the intrinsic value of progress of the scientific field from at least two perspectives. First, from a practitioner point of view, identifying detrimental factors for decision quality helps to optimize processes for human decision makers, and to facilitate selection of decision makers with highest expected performance. Second, from a societal point of view, the identification of influences on the quality of individual financial decisions might help isolating factors that place vulnerable people at a systematic disadvantage. This knowledge allows to devise interventions aimed at tackling those factors, and, thus, provide powerful levers to fight socioeconomic inequality.

Study 1: The effects of acute and chronic stress on choice consistency

In study 1a (Nitsch, Sellitto, et al., 2021b), we investigated the influence of acute social stress on choice consistency. Study 1b (Nitsch, Sellitto, et al., 2021a) provides a detailed description of the data and methods.

Stress can be defined as a state of fearful arousal in response to an environmental demand exceeding the perceived ability to cope (Fink, 2016). The necessary processing and appraisal of incoming sensory information is performed by limbic structures, including hippocampus, amygdala and prefrontal cortex (De Kloet et al., 2005). The stress response, then, recruits mostly two systems: The sympathetic-adrenal-medullary system and the hypothalamic-pituitary-adrenal axis.

The sympathetic-adrenal-medullary system activates immediately after stressor-onset. The locus coeruleus releases norepinephrine within the brain and activates the sympathetic nervous system. Sympathetic activation stimulates the secretion of catecholamines in the body: epinephrine (primarily from the adrenal medulla) and norepinephrine (directly from sympathetic nerves). This causes an increase of heart rate and blood pressure, peripheral vasoconstriction and energy mobilization (Ulrich-Lai & Herman, 2009). The catecholaminergic response normalizes within 30 to 60 minutes after stressor- onset (Hermans et al., 2014).

Simultaneously, the hypothalamic-pituitary-adrenal axis is activated: A cascade of hormonal secretion is triggered, including corticotropin-releasing hormone (CRF) from the paraventricular nucleus of the hypothalamus, adrenocorticotropin (ACTH) from the pituitary gland and cortisol from the adrenal cortex (Joëls & Baram, 2009). Cortisol affects the brain both through rapid, non-genomic and slow, genomic effects (Groeneweg et al., 2011). Rapid, non-genomic effects include a negative feedback-loop with the hypothalamic-pituitary-adrenal axis through inhibition of the hypothalamus (Evanson et al., 2010) and pituitary gland (Hinz & Hirschelmann, 2000) that normalizes cortisol levels medium-term. The cortisol level in the brain peaks around 20 minutes after stressor-onset (Droste et al., 2008) and normalizes within two hours after stressor-onset (De Kloet et al., 2005) with effects on the brain persisting for several hours.

An often-neglected aspect of the stress response is its dependent mode of operation. Its effective neuronal time-profile might best be described by a framework of three temporal domains (Joëls & Baram, 2009). Immediately after stress, norepinephrine upregulates the salience-network including amygdala, dorsal anterior cingulate cortex, anterior insula, thalamus, inferotemporal/temporoparietal regions, striatum and brainstem (Hermans et al., 2011). Simultaneously, it downregulates the executive-control network including dorsolateral and medial prefrontal cortex, frontal eye fields and dorsal posterior parietal cortex, and is associated with an impaired prefrontal function (Arnsten, 2009; Arnsten et al., 2012; Qin et al., 2009). With slight delay after stress, rapid, non-genomic effects of cortisol facilitate or inhibit the transmission of ion channels, receptors and neurotransmitters within different limbic and brain stem structures (Tasker et al., 2006). They are thought to interact with and enhance catecholaminergic effects (Hermans et al., 2014). In the aftermath of stress, genomic cortisol effects take over, gradually, through altered gene-transcription by mineralocorticoid and glucocorticoid receptors. These genomic cortisol effects change structural integrity and excitability of the receptors. Mineralocorticoid receptors are responsible for maintaining stress-related neural circuits whereas glucocorticoid receptors are working to reestablish homeostasis (De Kloet et al., 2005). Glucocorticoid receptor-mediated effects several hours after stressor-onset enhance the function of the prefrontal cortex and hippocampus but inhibit function of the amygdala. They might provide a mechanism that actively reverses the rapid effects of catecholamines and cortisol (Hermans et al., 2014).

Given this temporally-dynamic theoretical framework of the acute stress response and previous findings regarding underlying cognitive decision-making abilities (Beilock & DeCaro, 2007; Brand et al., 2005, 2014; Frederick, 2005; Gathmann et al., 2014; Maier et al., 2015; Margittai et al., 2016; Qin et al., 2009), we hypothesized that acute stress affects rationality in a similarly time-dependent fashion. Further, we explored the potential influence of self-reported chronic stress. For detailed information of the used methodology see Methods of Study 1a, and Study 1b. Study 1a was published as a research article in *Psychoneuroendocrinology*, Study 1b was published in *Data in Brief*.

Study 2: Influence of memory processes on choice consistency

In study 2 (Nitsch & Kalenscher, 2021a), we investigated the influence of memory processes on choice consistency. Previous theoretical and empirical research suggests an important role of memory processes in decision-making, for example for the construction of goals and preferences (Gabaix & Laibson, 2017; Johnson et al., 2007; Wimmer & Shohamy, 2012). However, memory processes (and their failure) are not directly observable in choice behavior and hardly manipulable for value-based choice, which poses a challenge for classic revealed preference methodology. To overcome this challenge, we generalized GARP to the domain of perceptual decisions (Nitsch & Kalenscher, 2021a, pp. 12–14) and developed a novel multi-attribute visual choice task (Nitsch & Kalenscher, 2021a, pp. 5–6, 9–11). This allowed us to experimentally manipulate the representation strength of choice goals via the variation of the retention interval for a visual exemplar. As choice consistency requires a clearly defined structure of goals (Afriat, 1973), we hypothesized that an increased retention interval leads to lower visual choice consistency. The effectiveness of the memory manipulation was confirmed in a pilot experiment and in a manipulation check. For detailed information of the used methodology see Methods of Study 2. The whole study was performed as a Registered Report, and is accepted for publication at *Royal Society Open Science*.

Study 3 and 4: On the robustness of choice consistency

Next to our own experimental work, we further conducted two analyses of the literature body to answer the question of how robust choice consistency as a proxy for rationality is to the influence of internal and external factors, generally.

Specifically, in study 3 (Nitsch & Kalenscher, 2020) we used an unsystematic, qualitative approach to review publications on influence factors of choice consistency strictly operationalized via revealed preference theory. Due to the relative novelty of the research agenda and, therefore, the still limited amount of studies published, we were able to qualitatively evaluate each study identified in our search. However, there were also two obvious limitations to the approach: First, the limited amount of studies considered often prohibited general conclusions regarding specific influence factors. Second, an unsystematic and qualitative approach bears a high risk of bias. For detailed information see Study 3.

To address the concerns above, in study 4 (Nitsch & Kalenscher, 2021b) we conducted a systematic, quantitative review of the literature body. Further, we broadened our scope to accept various operationalizations of choice consistency (see Study 3, Supplemental Material). Finally, to address the concern of bias we used P-Curve analysis (Simonsohn et al., 2014a, 2014b, 2015).

P-Curve analysis uses the distribution of p-values in the published literature to make inferences about the presence of a true effect. Due to the risk of publication bias against non-significant results, only p-values smaller than 0.05 are considered. The null hypothesis of no true effect in the literature is then characterized by a flat distribution of p-values in that range, whereas the presence of an effect would result in a right-skewed distribution. Various statistical tests can be performed to test the null hypothesis, some of which are even robust to more ambitious fraudulent statistical practices (Simonsohn et al., 2015). For detailed information of the used methodology see Supplemental Methods of Study 4. Both manuscripts have been published as preprints. Study 4 was accepted as a virtual posted at the annual meeting of the Society for Neuroeconomics.

Study 5: On the reliability of choice consistency

Many studies of the contemporary research program on influence factors of rationality have practically treated choice consistency as a psychometric measure. And, indeed, based on its rich theoretical (and philosophical) foundations, choice consistency in the sense of revealed preference arguably has good validity as a measure. However, a second desirable property of psychometric measures according to classical test theory, reliability, has been neglected so far. Even worse, previous theoretical work on the reliability of behavioral tasks suggests that tasks with low between-subject variance tend to show low reliability (Hedge et al., 2018). Given that many influential studies deployed correlational designs (e.g. Choi et al., 2014; Chung et al., 2017; Kim et al., 2018) that are particularly at risk to low reliability, in Study 5 (Nitsch, Lüpken, et al., 2021) we set out to determine the test-retest- and inter-method-reliability of revealed preference choice consistency in the domain of social decisions. For detailed information of the used methodology see Methods of Study 5. Study 5 has been published as preprint.

Studies

Study 1a – The effects of acute and chronic stress on choice consistency

Nitsch, F. J.*, Sellitto, M., & Kalenscher, T. (2021). The effects of acute and chronic stress on choice consistency. *Psychoneuroendocrinology*, *131*, 105289. https://doi.org/10/gk9pvk

*Corresponding Author

CRediT Author Statement:

Felix Jan Nitsch: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Original draft preparation, Visualization, Project Administration

Manuela Sellitto: Data curation, Writing - Review & Editing, Supervision

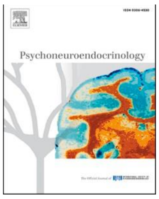
Tobias Kalenscher: Conceptualization, Methodology, Resources, Writing - Review & Editing, Supervision, Funding Acquisition

Psychoneuroendocrinology 131 (2021) 105289

Contents lists available at ScienceDirect

Psychoneuroendocrinology

journal homepage: www.elsevier.com/locate/psyneuen



The effects of acute and chronic stress on choice consistency[★,★★]

Felix J. Nitsch [*,1], Manuela Sellitto [2], Tobias Kalenscher [3]

Comparative Psychology, Heinrich-Heine-University Düsseldorf, Germany

ARTICLE INFO

Keywords:
Stress
Choice consistency
Rationality
Cortisol
Revealed preference

ABSTRACT

Important decisions are often made under some degree of stress. It is now well-established that acute stress affects preferences and priorities in our decisions. However, it is hard to make a general case on the net impact of stress on decision-making quality in a normative sense as evidence for or against a direct effect of stress on decision-making quality is sparse. Here, we used the revealed preference framework of choice consistency to investigate decision-making quality without the assumption of an objectively correct choice. Specifically, we tested whether acute stress influences choice consistency in a time dependent fashion. A sample of 144 participants solved a food choice task before, immediately after and in the aftermath of the Trier Social Stress Test (TSST) or a matched control procedure. We confirmed the effectiveness of our stress manipulation via an array of subjective and physiological stress measures. Using Bayesian statistics, we found strong evidence against an effect of acute stress on choice consistency. However, we found exploratory evidence for a negative association of self-reported chronic stress and choice consistency. We discuss our results in the context of previous findings of stress effects on choice consistency and preference changes.

1. . Introduction

Important decisions are often made under some degree of stress. In an increasingly fast-paced economy, managers have to decide whether to keep or kill a project with high financial and personal impact. Doctors have to select the most promising therapy when lives are at stake. Air traffic controllers have to declare a flight safe to start, evaluating a complex multi-factorial system. It is crucial to understand how stress might affect the ability to make good decisions.

In modern economics, decision-making quality is not defined by the specific direction of preferences but instead by bounds on the preference structure in the form of consistency principles (Sugden, 1991). Revealed preference theory (Houthakker, 1950; Samuelson, 1938; Varian, 1982) requires decision-makers to have a well-defined and stable set of preferences, as well as to act cost-efficiently and consistently upon their fulfillment, which is mathematically equivalent with optimizing decision-making (Afriat, 1973). Importantly, revealed preference optimality does not require objectively, or at least normatively correct preferences (Choi et al., 2014). The notion of decision-making quality as choice consistency can be illustrated by an example:

Assume that a manager chooses to endorse a project A instead of another project B when both require the same budget. The same manager must, everything else equal, also endorse project A when it is cheaper than project B. That is, because from the former decision we learn that the manager either prefers project A or is indifferent and chose randomly (because both projects were equally expensive). Now the necessity of choosing project A in the latter decision can be proven by exhaustion: If the manager prefers project A over B, they should always choose project A as long as it is at least as cheap as project B. If the manager is indifferent between both projects, they should always choose the cheaper one, which is project A in the second decision, in order to act cost-efficiently. Also, transitivity must hold for these *revealed preferences. For example, if the manager chose project A over B and project B over C, always assuming equal project costs, then they should not choose project C*

[★] The research was supported by a grant from the Deutsche Forschungsgemeinschaft (Bonn, Germany) to Tobias Kalenscher (grant-ID: DFG-KA 2675/4-3). [★★] We thank Adam Schweda for helpful comments on the study design as well as assistance with the data collection. We thank Johanna Falk, Anne Schmittner, Jochen Hoog, Alina Keßler and Paul Kramer for assistance with the data collection.

[*] Corresponding author.
E-mail address: felix.nitsch@hhu.de (F.J. Nitsch).

[1] https://orcid.org/0000-0002-7832-7498.

[2] https://orcid.org/0000-0002-3987-7526.

[3] https://orcid.org/0000-0002-0358-9020.

https://doi.org/10.1016/j.psyneuen.2021.105289
Received 27 January 2021; Received in revised form 18 May 2021; Accepted 25 May 2021
Available online 31 May 2021

over project A, when project A is cheaper. In our example, it is irrelevant whether and why the manager prefers project A or B or even is indifferent, as long as they are, everything else equal, consistent in their decisions.

Among several other factors, stress has been shown to affect the preferences and priorities in our decisions concerning food (Habhab et al., 2009; Maier et al., 2015; Zellner et al., 2006), other people (Margittai et al., 2015; Margittai, Van Wingerden, et al., 2018; Schweda et al., 2019; Vinkers et al., 2013; Von Dawans et al., 2012), risky prospects (Margittai, Nave, et al., 2018; Starcke and Brand, 2016), as well as the timing of financial returns (Cornelisse et al., 2013; Kimura et al., 2013; Riis-Vestergaard et al., 2018). These behavioral effects are paralleled by resource reallocation across large-scale brain networks during the stress response (Hermans et al., 2011, 2014). Immediately after the stressful event, catecholamines upregulate vigilance-related functions while prefrontal functions including executive control are inhibited (Arnsten, 2009; Arnsten et al., 2012; Hermans et al., 2011; Qin et al., 2009). With slight temporal offset, this pattern is enhanced by rapid, non-genomic corticosteroid effects (Hermans et al., 2011; Tasker et al., 2006). In the aftermath of stress genomic, corticosteroid effects are working to reestablish homeostasis (De Kloet et al., 2005), but repeated stress exposure slows the recovery of homeostasis after stressful events (Cameron and Schoenfeld, 2019).

Hence, stress affects preferences in various choice domains. However, as argued above, such preference shifts are not sufficient to show decreased decision quality. For example, the increased preference for calorie-dense food and immediate rewards under stress might simply reflect the attempt to replenish depleted energy resources; hence, choices for high-caloric food might very well be consistent with the decision-maker's preferences.

A more convincing argument for the possibility that stress could impair decision-making quality stems from the observation that underlying cognitive decision-making abilities deteriorate under stress. For example, acute stress, or stress hormone action, has been shown to impair working-memory, problem-solving abilities, self-control, and cognitive reflection (Beilock and DeCaro, 2007; Maier et al., 2015; Margittai et al., 2016; Qin et al., 2009). All of these psychological functions have been linked to decision-making (Brand et al., 2005, 2014; Frederick, 2005; Gathmann et al., 2014). In addition, repeated exposure to stress (i.e. chronic stress) has been robustly linked to decreased mental health, particularly depression, an increased number of cognitive failures as well as neurotoxic effects in general (Cameron and Schoenfeld, 2019; Linden et al., 2005; Marin et al., 2011). Animal research has provided evidence that chronic stress affects the ability to perform actions based on their consequences (Dias-Ferreira et al., 2009), which is an integral feature of utilitarian decision-making. Further, chronic stress in animals has been linked to increased high-cost/high-reward choices over a wide range of options (Friedman et al., 2017). Still, it does not follow immediately, that decision-making quality must be impaired if certain cognitive abilities are. Through means of resource reallocation, deficiencies in one ability might be compensated by an upsurge in another.

Overall, evidence for or against a direct effect of stress on decision-making quality is sparse, with only one study testing the influence of an acute physical stressor on choice consistency in decisions among risky prospects (Cettolin et al., 2019). Here, we asked whether acute stress impairs decision-making quality in the revealed preference sense defined above. That is, we asked whether acute stress increases the tendency to make decisions against one's previously revealed preferences. Previous research suggests that the immediate effects of the stress response impairs functions important for decision-making such as executive control, while the aftermath of the stress response enhances them. Therefore, we tested the following hypothesis:

Hypothesis: Acute stress influences choice consistency in a time dependent fashion similar to the effects on executive functions. Specifically, we expect choice consistency to be impaired immediately after stressor offset, but not in the aftermath of a stressor.

Further, cognitive functioning is impaired under chronic stress exposure. While the investigation of chronic stress was not the main objective of the current study, we assessed self-reported chronic stress exposure to explore its influence on choice consistency.

Explorative question: Is chronic stress exposure associated with impaired choice consistency?

2. Methods

2.1. Sample characteristics and inclusion criteria

We included 144 participants (76 females). Participants did not have formal psychological or economic education, were 18 – 40 years old, non-smokers and did not take medication that could have influenced their corticosteroid levels. Women were not taking oral contraceptives. Similar to previous studies (Margittai et al., 2015; Schweda et al., 2019), participants had to refrain from drinking alcohol and sexual activities 24 h, caffeine four hours and eating/drinking (except water) two hours before the experiment. Table 1 provides summary statistics of the sample characteristics.

Table 1
Demographic and trait measures data per experimental group.

Demographic Variable	Experimental group	Control group	BF_{10} (for group diff.)
N	N = 75	N = 69	
Age	$M = 24.10, SD = 5.14$	$M = 24.80, SD = 4.48$	0.27
Sex	*female*:39, *male*:36	*female*:37, *male*:32	0.21
University Degree	*yes*:52, *no*:23	*yes*:50, *no*:19	0.19
TICS	$M = 140.00, SD = 27.00$	$M = 149.00, SD = 24.20$	1.11
BFI-10			
Extraversion	$M = 6.33, SD = 1.95$	$M = 6.49, SD = 1.75$	0.20
Agreeableness	$M = 6.65, SD = 1.78$	$M = 6.35, SD = 1.68$	0.30
Conscientiousness	$M = 6.76, SD = 1.63$	$M = 6.52, SD = 1.67$	0.25
Neuroticism	$M = 5.89, SD = 1.98$	$M = 6.06, SD = 1.92$	0.20
Openness	$M = 7.29, SD = 1.89$	$M = 7.71, SD = 1.78$	0.42
BIS/BAS			
Behavioral Inhibition	$M = 15.00, SD = 3.85$	$M = 14.10, SD = 3.61$	0.52
Drive	$M = 7.76, SD = 2.30$	$M = 7.51, SD = 1.84$	0.23
Fun Seeking	$M = 7.64, SD = 1.63$	$M = 7.36, SD = 1.67$	0.29
Reward Responsiveness	$M = 8.20, SD = 2.32$	$M = 7.87, SD = 1.98$	0.26
QDQ	$M = 36.10, SD = 5.78$	$M = 33.80, SD = 5.53$	2.29
SDS	$M = 21.40, SD = 3.00$	$M = 21.00, SD = 2.86$	0.23
MEQ	$M = 15.90, SD = 1.75$	$M = 15.80, SD = 2.04$	0.18
MWT-B	$M = 27.20, SD = 4.18$	$M = 28.20, SD = 4.18$	0.46

Bayes factors were calculated using *t*-tests for continuous dependent variables and contingency tables for categorical dependent variables with default priors. Abbreviations: TICS = Trier Inventory of Chronic Stress, BFI-10 = 10-item version of the Big Five Inventory, BIS/BAS = Behavioral Inhibition/Activation Scale, QDQ = Quick Delay questionnaire, SDS = Social Desirability Scale, MEQ = Morningness-Eveningness-Scale, MWT-B = multiple-choice vocabulary test (German: Multipler Wortschatz Test).

2.2. Procedure

Before the experimental session, all participants were screened for the inclusion criteria, rated their liking of several fruits/vegetables and sweet/salty snacks on a 5-point Likert scale online, and responded to various psychological trait measures via an online questionnaire (see trait measures).

All experimental sessions took place from 3 p.m. to 6 p.m. to control for circadian variations of hormonal levels. Participants were assigned to the two experimental conditions pseudo-randomly. In the lab, participants completed the behavioral task at three different time points (Fig. 1): on arrival (*baseline time point*), within ten minutes after the stress/control procedure offset (*early time point*), and 90 min after stress/control procedure offset (*late time point*). Between the early and the late time point, participants were allowed to read magazines but not to talk or to use their phones.

2.3. Stress manipulation

In order to experimentally induce acute psychosocial stress, we used the group version of the Trier Social Stress Test (TSST-G): During the 20 min long TSST-G procedure, groups of four women or men were asked to carry out a fictional job interview (3 min per participant) and a mental arithmetic task (deducting from a four-digit number in steps of 16; 1.5 min per participant) in front of an evaluative panel of experts while being videotaped. The panel consisted of a female and a male panelist. For women, the male panelist took up the active role and the female panelist, vice-versa, for men. Panelist wore white lab coats. They closely observed the participants, interrupted them and took notes of their performance, and did not give any verbal or non-verbal positive feedback. We used a matched control procedure for the control group (von Dawans et al., 2011), in which groups of three to four women or men were instructed to tell a story about a good friend and complete the mental arithmetic task simultaneously. To compensate for the shorter, simultaneous procedure, participants were instructed to read magazines in between tasks. They were neither videotaped nor directly observed by the panel, removing the component of social evaluation but keeping other aspects of the procedure as similar as possible. The panel also did not were lab coats in the control condition. The TSST is considered the gold-standard for stress-induction in humans and has been validated extensively (Het et al., 2009; Kirschbaum et al., 1993; Kudielka and Kirschbaum, 2005; Rohleder et al., 2004) and successfully established in our lab (Margittai et al., 2015; Schweda et al., 2019).

2.4. Physiological and subjective stress measures

To assess the effectiveness of the stress manipulation, we collected a saliva sample before and after each behavioral task time point as well as during the stress/control procedure and measured cortisol and alpha-amylase, a measure of noradrenergic activity (Nater & Rohleder, 2009), for a total of 7 measurements across the experiment (Fig. 1, upper panel). Saliva samples were collected using Salivette devices (Sarstedt, Germany) containing a cotton wool swab that participants had to lightly chew on for 60 s to allow the swab to fill with saliva. Samples were frozen and stored at −20 degree Celsius and subsequently sent to the Dresden LabService GmbH for cortisol and alpha-amylase measurement. Both cortisol and alpha-amylase concentrations were determined using a luminescent immunoassay (IBL International, Hamburg). For 16 samples analysis of cortisol and/or alpha-amylase was not possible due to insufficient saliva or contamination with blood. Furthermore, we assessed the heartrate of the participants at baseline and during the stress/control procedure with commercial wristbands (Polar A370) and their current affect with the Positive and Negative Affect Scale (Krohne et al., 1996) and four visual analogue scales (items: *stressed*, *ashamed*, *insecure*, *self-secure;* 100 mm scale). We assessed the natural chronic stress exposure of our participants prior to the experiment with the Trier Inventory of Chronic Stress (TICS; Petrowski et al., 2012). The questionnaire comprises 57 items, which ask for 9 types of stress. These include work overload, social overload, pressure to succeed, dissatisfaction with the job, excessive demands at the job, lack of social recognition, social tension, social isolation and chronic worries. Each item must be scored on a five-level scale from 0 to 4. These values are added together to form the total score. A higher score corresponds to higher chronic stress exposure.

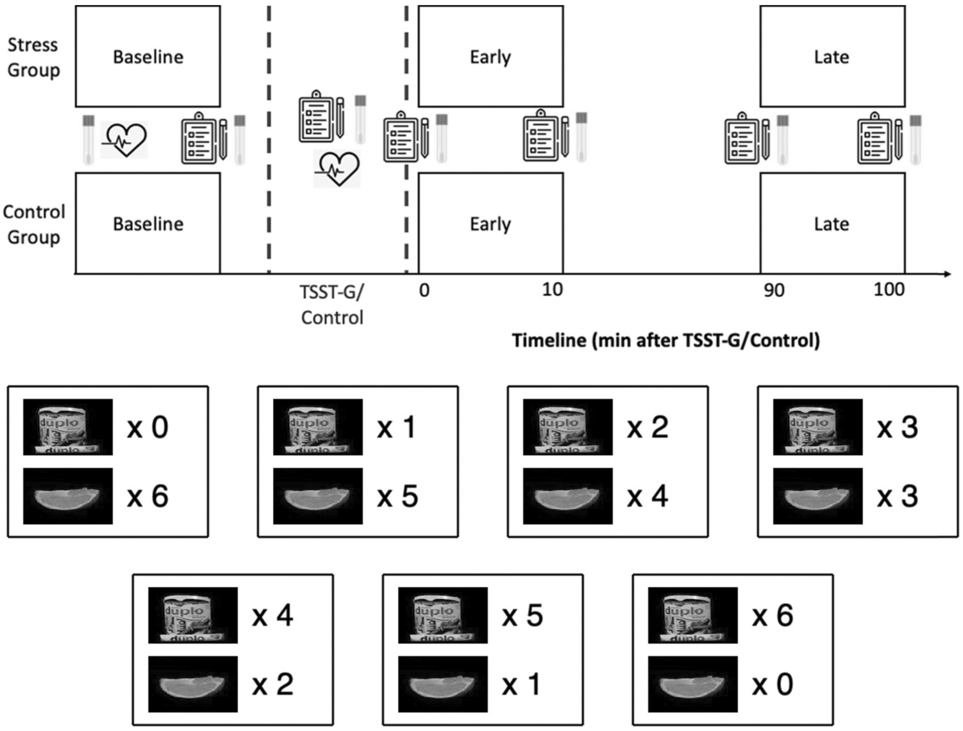


Fig. 1. Experimental timeline, measurements and task Upper panel. Experimental timeline including measurements. Saliva samples were taken before and after each completion of the food choice task as well as during the TSST-G/control procedure. Mean heartrate was measured during baseline and during the TSST-G/control procedure. The self-report measures were taken before and during the TSST-G/control procedure, as well as before and after every subsequent completion of the food choice task. The time on the x-axis is not true to scale. Lower panel. Example trial of the food choice task. In this particular trial, the choice set consists of 7 different snack bundles ranging from 6 orange pieces and 0 chocolate bars to 0 orange pieces and 6 chocolate bars. A choice set is always given by all choice bundles whose cost fully exhausts but not overspends the budget given the current prices of both snack types. By increasing and decreasing the prices as well as varying their ratio, we receive the 22 different choice sets of the experiment. For the example trial, chocolate bars and orange pieces were equally expensive, allowing to purchase exactly 6 units of snacks. Importantly, participants were only presented with the final choice set, but not prices or budgets to minimize the difficulty of understanding the task.

2.5. Trait measures

Besides the TICS, we administered the following questionnaires prior to the experiment to exclude the possibility of spurious trait differences between stress and control participants: the 10-item version of the Big-Five-Inventory (BFI-10; Rammstedt and John, 2007), the behavioral inhibition & activation scale (BIS/BAS; Strobel et al., 2001), the quick delay questionnaire (QDQ; Clare et al., 2010), the social desirability scale (SDS; Stöber, 1999) and the morningness-eveningness-scale (MEQ; Griefahn et al., 2001). Furthermore, we assessed verbal intelligence via the multiple-choice vocabulary test (MWT-B; Lehrl et al., 1995) at the end of the laboratory session, as participants could have easily cheated in an online version.

2.6. Food choice task

We deployed a standard food choice task similar to the one used by Harbaugh et al. (2001) and Chung et al. (2017). We specifically decided for a food choice task over other choice domains (i.e. risk or time preferences) to minimize task complexity and, thus, ensure that any inconsistencies would not arise simply from misunderstanding of the task. This was especially important as we recruited a sample that was naïve to economic theory.

In each trial, participants had to choose one out of a set of two to seven snack bundles. Each snack bundle consisted of specific amounts of a sweet or salty snack and a fruit or vegetable (see Fig. 1, lower panel). The snacks for both categories were selected to be similarly attractive according to the previously provided online ratings of the participants. At each of the three time points, participants had to make choices in 11 trials. The 11 trials for each time point were randomly sampled from a collection of 22 possible trials (see Supplemental Fig. S1). The sampling procedure was implemented to reduce interdependency of the answers of each participant for subsequent time points, while keeping the presented bundle size out of satiation range. At the end of the experiment one random decision of each participant was implemented and the participant received the corresponding snack bundle. Using simulated data of 10.000 uniform-randomly deciding virtual decision-makers, we ensured that our design was sufficiently powerful to detect inconsistent behavior (see below for a quantitative definition of inconsistency): $Bronar's\ Power = 0.972$ (Bronars, 1987). Supplemental Fig. S2 summarizes the results of our simulation study.

For 8 participants, no or incomplete food choice data were saved due to a technical failure of the experimental hardware.

2.7. Analysis pipeline

For all analyses, we used a Bayesian framework of inference. Bayesian statistics allows us to express confidence in that a parameter is within a certain range, to extend parameter estimation naturally for complicated models, and to express evidence for or against hypotheses on a continuous scale (Wagenmakers et al., 2018). The latter point is especially relevant if we want to make statements about the absence of a possible effect, not only the absence of evidence. Raw and processed data, study materials, as well as the analysis code to computationally reproduce our results are available online and publicy: https://osf.io/6mvq7/

2.7.1. Decision-making quality analysis

Instead of a normative evaluation of preferences, we defined decision quality as the consistency with which participants pursued their individual preferences regardless of their direction. That is, we assessed choice consistency, not choice content. Notably, this implies that choices of high-caloric, high-fat and high-sugar foods could potentially be considered rational, as long as such choices were consistent.

To this means, we calculated the Critical Cost Efficiency Index (Afriat, 1972; Varian, 1991) for each participant per time point. The Critical Cost Efficiency Index (CCEI) is the most widely used measure of rationality or decision-making quality in the framework of revealed preference theory. The CCEI exploits the fact that, in theory, inconsistent behavior is not cost-efficient, which can be illustrated by an example:

Assume, that from a previous choice we learn that 5 chocolate bars are at least as valuable to a decision-maker as 5 pieces of orange. Let's further assume that at another point in time, the decision-maker still buys 5 orange pieces although they cost 10% more than 5 chocolate bars. As we know that 5 chocolate bars are at least as good as 5 orange pieces, they spent at least 10% too much money for the obtain valued and, conversely, showed a 90% cost efficiency.

This would be denoted by a CCEI of 0.9. A CCEI of 1 denotes 100% cost efficiency and perfect consistency. The index approaches zero as the behavior becomes more inconsistent (for a different interpretation of the CCEI see also Nitsch & Kalenscher, 2020).

2.7.2. Hormonal analysis

Values were not subjected to any data transformations (Feng et al., 2014). We excluded one data point from our cortisol data (control group, measurement "TSST-G/control", VPN111) which deviated more than 20 standard deviations from the grand mean of our cohort and was more than 60 times higher than the median reference value for females her age, controlling for wake-up time (Miller et al., 2016).

2.7.3. Statistical model

We used a mixed-factorial design with time point (baseline, early, late) as a within-subject factor and experimental group (stress, control) as a between-subject factor. In statistical terms we used a mixed-factor ANOVA-style model including time point as within-subject factor and experimental group as between-subject factor, and the CCEI as dependent variable. Our main interest lies in the possible interaction of both factors as a statistical representation of the time-dependent influence of acute stress on choice consistency (see hypotheses). Hence, our interpretation of the data will be guided by two model comparisons:

1. The evidence for the full model including both main effects as well as the interaction effect compared to a null model (including only a subject-level intercept).
2. The evidence for the full model including both main effects as well as the interaction effect compared to a reduced model only containing both main effects but not the interaction effect.

As previous evidence on the time-dependent effect of acute stress on choice consistency is limited, we will use uninformative default priors for our model parameters (Rouder et al., 2012). We used structurally similar models for our manipulation checks (physiological and subjective stress measures). To test for an association of self-reported chronic stress and choice consistency we used a Bayes Factor test for correlations. Lastly, to test for absence of trait differences between stress and control participants we used pair-wise Bayesian *t*-tests contingency tables for the assessed questionnaire scores and demographic variables.

2.8. Data collection plan

Our data collection plan was based on a Bayesian stopping rule: We planned to collect data until we had reached a Bayes factor of $BF \geq 10$ V $BF \leq 0.1$ for all manipulation checks as well as our hypothesis, but at least for one year, to ensure interpretability of our analysis. Evidence for our explorative question was secondary to our sample size rationale and explorative results should be interpreted accordingly.

3. Results

3.1. Demographic variables and trait measures

We did not find strong evidence for any difference of the two experimental groups regarding all demographic variables and trait measures considered (see Table 1). Accordingly, we can consider our randomization of participants into experimental conditions successful.

3.2. Cortisol and alpha-amylase

The data for both hormonal markers indicated that our stress manipulation was successful. The full model including both main effects and the interaction effect was much more likely than the null model given the data for both cortisol ($BF_{10} > 100$) and amylase ($BF_{10} > 100$), which is conventionally interpreted as extreme evidence (Jeffreys, 1998). Similarly, we found extreme evidence for the inclusion of the interaction effect by comparing the full model to reduced model for both cortisol ($BF_{10} > 100$) and amylase ($BF_{10} > 100$). Fig. 2 shows a plot of the data. In summary, we found extreme evidence for the effectiveness of our stress manipulation in increasing the levels of salivary cortisol and alpha-amylase.

3.3. Heartrate

The effectiveness of our stress manipulation was corroborated by the heartrate data. The full model including both main effects and the interaction effect was much more likely than the null model given the data ($BF_{10} > 100$), which is conventionally interpreted as extreme evidence (Jeffreys, 1998). Similarly, we found extreme evidence for the inclusion of the interaction effect by comparing the full model to reduced model ($BF_{incl} > 100$). Fig. 3 shows a plot of the data. In summary, we found extreme evidence for the effectiveness of our stress manipulation as operationalized by the heartrate.

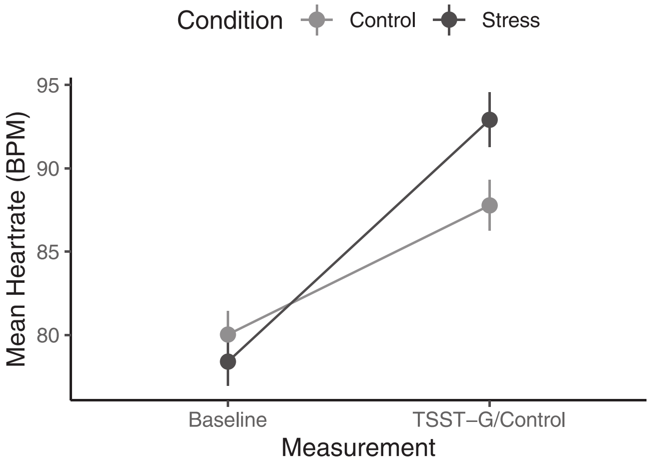


Fig. 3. Manipulation check: Mean heartrate. Depicted are the mean heartrate (in beats per minute) at baseline and during the TSST-G/control procedure. Error bars show the standard error of the mean.

3.4. Self-reported affect

Lastly, stress dynamically influenced positive and negative affect, as well as self-reported stress, shame and insecurity (but not security): We found strong or extreme evidence for the effectiveness of our stress manipulation as operationalized by 5 out of 6 of the self-report scales (see supplemental table). Fig. 4 shows a plot of the data.

Hypothesis: Acute stress and choice consistency.

Our behavioral results indicated that participants acutely stressed by the TSST-G did not dynamically differ in their choice consistency

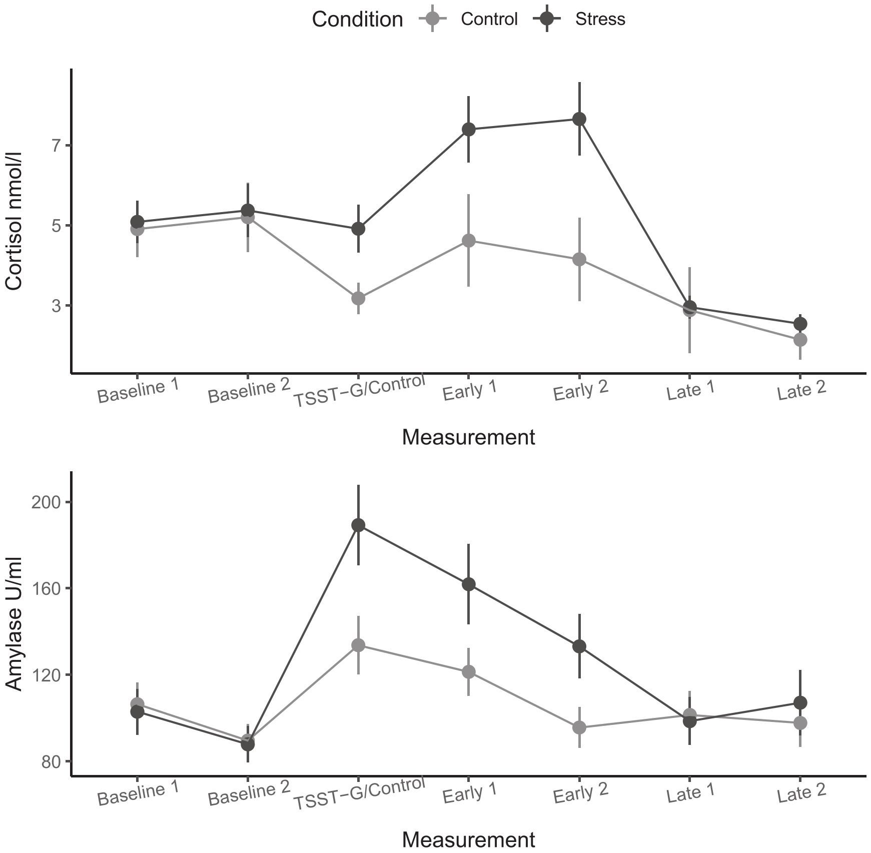


Fig. 2. Manipulation check: Cortisol & alpha-amylase levels. Depicted are the mean cortisol (top) and alpha amylase (bottom) levels across the experimental timeline. Error bars show the standard error of the mean. The sample time-points depicted on the x-axis refer to the pre- and post-TSST-G samples, as shown in Fig. 1. For cortisol, group differences peaked after the first behavioral test and decayed over the course of the experiment. For alpha amylase, group differences peaked during the TSST-G/ control procedure and decayed over the course of the experiment.

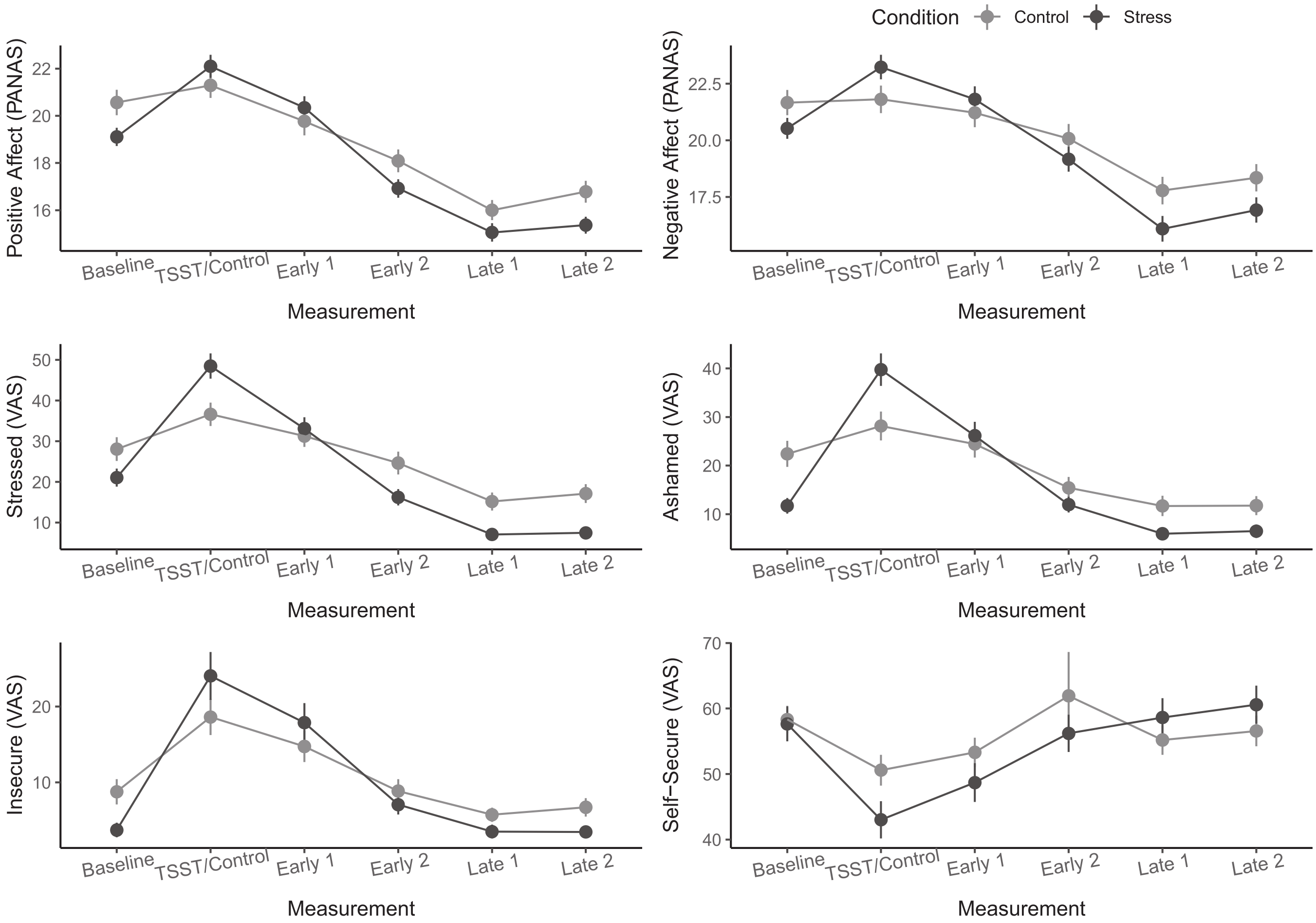


Fig. 4. Manipulation check: Self-reported affect. Depicted are the mean response levels for all self-reported affect measures across the experimental timeline. Error bars show the standard error of the mean. Generally, group differences for self-reported affect peaked during the TSST-G/control procedure and decayed or even reversed over the course of the experiment. This reversal might be explained via arousal/fatigue dynamics.

compared to not-stressed participants. The null model was 24 times more likely than the full model given the data ($BF_{10} = 24.23 \pm 2.27\%$), which conventionally can be interpreted as strong evidence for the absence of an effect (Jeffreys, 1998). Similarly, we found strong evidence against the inclusion of the interaction effect by comparing the full model to a reduced model ($BF_{incl} = 15.26 \pm 4.74\%$; see Fig. 5, upper panel). To corroborate the robustness of our results, we repeated the analysis including self-reported sex and age as covariates into all of our models. Results remained qualitatively unchanged ($BF_{10} = 23.61 \pm 3.77\%$; $BF_{incl} = 14.92 \pm 4.19\%$). In summary, we found strong evidence against our hypothesis. That is, participants in the TSST-G (acute stress) condition did not dynamically differ in their choice consistency compared to not-stressed participants but, instead, showed comparable consistency levels across all time points of the experiment.

3.5. *Exploration: chronic stress and choice consistency*

We found exploratory evidence that choice consistency decreased with higher chronic stress exposure (see Fig. 5, lower panel). A negative relationship of chronic stress and choice consistency was 9 times more likely than a null relationship given the data ($BF = 9.03$), which conventionally can be interpreted as moderate evidence (Jeffreys, 1998). The posterior estimate of the association under the assumption that it is non-zero was $\rho = -0.21, CI_{95\%} = [-0.36, \quad -0.05]$. This, tentatively, suggests the level of consistency decreased with increasing levels of chronic stress.

4. Discussion

Evidence for or against a direct effect of stress on decision-making quality is sparse, with only one other study testing the influence of a physical stressor on choice consistency in decisions among risky prospects (Cettolin et al., 2019). Here, we conducted an experimental test of the influence of a well-established social stress protocol on choice consistency in a food choice task. We specifically decided for a food choice task over tasks in other choice domains (i.e. risk or time preferences) to minimize task complexity and, thus, ensure that any inconsistencies would not arise simply from misunderstanding of the task. We corroborated the effectiveness of our stress induction using multiple subjective and physiological stress measures. Results showed strong evidence against a temporally dynamic effect of acute stress on choice consistency: Both experimental groups showed comparable choice consistency levels over all time points of the experiment. We conclude that decision quality does not deteriorate under acute stress. Further, we explored the relationship of self-reported chronic stress and choice consistency. Interestingly, our results indicated that, tentatively, higher levels of self-reported chronic stress were associated with lower choice consistency. Hence, decision quality might be impaired with increasing levels of chronic stress. This explorational result should be tested more rigorously in a future, confirmatory study.

Our results are in line with the findings of Cettolin et al. (2019), who did not find a significant effect of an acute physical stressor on choice consistency both shortly after stressor offset and in the aftermath of the stressor. However, although Cettolin et al. (2019) used a sufficiently

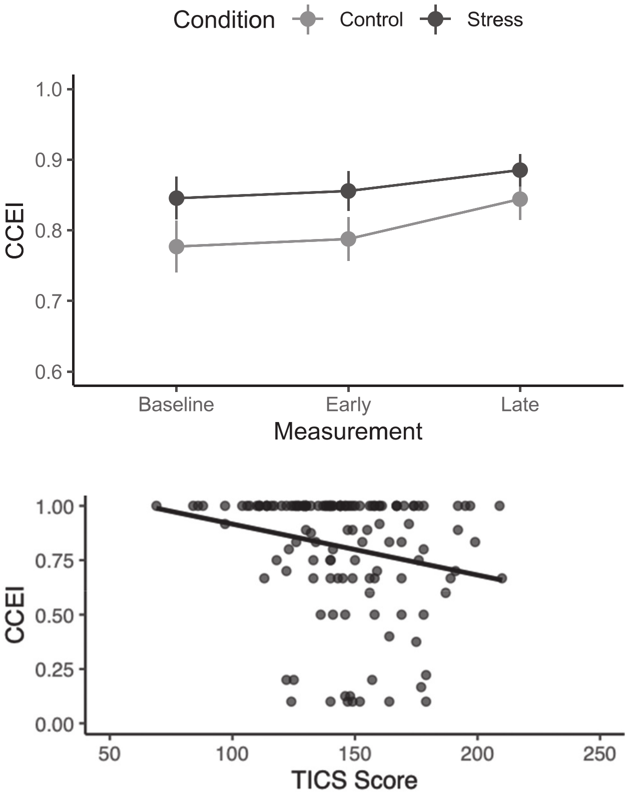


Fig. 5. Acute and chronic stress and choice consistency. Upper panel: Depicted are the mean choice consistency levels across the experimental timeline, measured by the Critical Cost Efficiency Index (CCEI). Error bars show the standard error of the mean. Choice consistency was mostly stable over the course of the experiment. Importantly, we found strong evidence against an interaction of time point and group. Note, that while descriptively there appears to be an overall difference between the two groups regarding choice consistency, there is no evidence favoring the inclusion of the experimental group factor compared to the null model ($BF_{10} = 0.76 \pm 0.89\%$). Lower panel; Each point in the scatter plot represents the data pair of the TICS score and the CCEI during the baseline measurement of a single subject. The trendline shows a least-squares regression. A negative relationship of chronic stress and choice consistency was 9 times more likely than a null relationship given the data (BF = 9.03). The empirical correlation coefficient was $\rho = -0.22$, the posterior estimate of the association under the assumption that it is non-zero was $\rho = -0.21$. A visual inspection of the scatter plot reveals that many participants are perfectly consistent. This is in line with previous studies (e.g. Choi et al., 2014).

large sample size and a statistically well-powered design, they did not explicitly quantify evidence for the null hypothesis. Using Bayesian statistics, we were able to provide a statistically well-founded confirmation of Cettolin and colleagues' conclusions. Further, we used a variant of the TSST as our stress protocol, which is generally considered the gold-standard for stress-induction in humans and has been validated extensively (Het et al., 2009; Kirschbaum et al., 1993; Kudielka & Kirschbaum, 2005; Rohleder et al., 2004). Our results showed that the null findings of Cettolin et al. (2019) generalize to a stress protocol with a social evaluation component, which is often regarded essential for the human stress response. Extending the findings of Cettolin and colleagues, we find explorative evidence that chronic stress, on the other hand, does affect choice consistency.

At first glance, these results seem to contradict previous results of dynamic preference shifts in response to stress, especially in the domain of dietary decisions (Habhab et al., 2009; Maier et al., 2015; Zellner et al., 2006). As stated above, choice consistency in the sense of revealed preference theory requires decision-makers to have a well-defined and stable set of preferences (Afriat, 1973). However, with our experimental procedure we ensured that each run of the food choice task for itself was completed within what we assume to be distinct time windows of the stress response (see Fig. 1). Thus, while we cannot exclude that preferences change between these time windows, we found strong evidence that choice consistency of stressed participants is not affected within these time windows as compared to not stressed participants. This is in so far intuitive, as altered preferences in response to a stressful event may very well be adaptive depending on the decision context, while a decreased level of cost-efficiency and choice consistency is rarely advantageous. Future research should challenge our results by investigating decision-making across hormonal time windows, e.g. by explicitly incorporating temporal dynamics in the modelling approach via leave-one-out measures of choice consistency (Kurtz-David et al., 2019).

It is important to point out that in the current study, our main focus was the effect of an acute stressful event on choice consistency in an incentivized choice task. While it is a potentially reassuring result that choice consistency is not impaired immediately following acute stress, or in its aftermath, we find tentative evidence that chronic stress does affect choice consistency. In line with this, animal research has found evidence that chronic stress affects the ability to perform actions based on their consequences (Dias-Ferreira et al., 2009), which is an integral feature of utilitarian decision-making. Another recent animal study found that chronic stress led to increased high-cost/high-reward choices over a wide range of options (Friedman et al., 2017). While the study of Friedman and colleagues could not determine whether this effect was driven by a reduced sensitivity to consequences, the authors argue that the cost-benefit integration was impaired in chronically stressed animals. Most likely due to methodological constraints in the investigation of chronic stress, evidence in humans is sparser. One study found that chronic stress in jockeys was related to impaired decisions in relatively simple attention and reaction time tasks (Landolt et al., 2017). The authors note that impairments, descriptively, increased with task complexity. More research is necessary to confirm whether chronic stress, in contrast to acute stress, impairs choice consistency, possibly via a reduced sensitivity to action outcomes. To this means, for example, the contemporary COVID-19 pandemic provides a unique opportunity to investigate the influence of increased chronic stress exposure on choice consistency in the general population.

A last concern that we want to address here is generalizability regarding the presentational structure of the choice problem and, related to that, the conceptualization of choice consistency. The investigation of choice consistency has increasingly diversified over the last decades and there exist several, prima facie equally valid operationalizations of choice consistency or, in a wider sense, rationality (Rieskamp et al., 2006). The here utilized revealed preference framework of choice consistency is a necessary and sufficient condition for as-if utility maximization (Afriat, 1973) and, thus, does have theoretical appeal. However, future research of choice consistency should address whether violations of choice consistency are specific to certain axioms, e.g. stochastic dominance, transitivity or regularity. Previous, seemingly contradictory findings in the field of choice consistency might be reconciled when using a stringent theoretical framework of choice consistency.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.psyneuen.2021.105289.

References

Afriat, S.N., 1972. Efficiency estimation of production functions (JSTOR). Int. Econ. Rev. 13 (3), 568–598. https://doi.org/10.2307/2525845.

Afriat, S.N., 1973. On a system of inequalities in demand analysis: an extension of the classical method. Int. Econ. Rev. 14, 460–472. https://doi.org/10.2307/2525934.

Arnsten, A.F., 2009. Stress signalling pathways that impair prefrontal cortex structure and function. Nat. Rev. Neurosci. 10 (6), 410–422.

Arnsten, A.F., Wang, M.J., Paspalas, C.D., 2012. Neuromodulation of thought: flexibilities and vulnerabilities in prefrontal cortical network synapses. Neuron 76 (1), 223–239.

Beilock, S.L., DeCaro, M.S., 2007. From poor performance to success under stress: working memory, strategy selection, and mathematical problem solving under pressure. J. Exp. Psychol.: Learn. Mem. Cogn. 33 (6), 983–998. https://doi.org/10.1037/0278-7393.33.6.983.

Brand, M., Fujiwara, E., Borsutzky, S., Kalbe, E., Kessler, J., Markowitsch, H.J., 2005. Decision-making deficits of Korsakoff patients in a new gambling task with explicit rules: associations with executive functions. Neuropsychology 19 (3), 267–277. https://doi.org/10.1037/0894-4105.19.3.267.

Brand, M., Schiebener, J., Pertl, M.-T., Delazer, M., 2014. Know the risk, take the win: how executive functions and probability processing influence advantageous decision making under risk conditions (Scopus). J. Clin. Exp. Neuropsychol. 36 (9), 914–929. https://doi.org/10.1080/13803395.2014.955783.

Bronars, S.G., 1987. The power of nonparametric tests of preference maximization. Econometrica 55, 693–698. https://doi.org/10.2307/1913608.

Cameron, H.A., Schoenfeld, T.J., 2019. Behav. Struct. Adapt. Stress 21.

Cettolin, E., Dalton, P.S., Kop, W.J., Zhang, W., 2019. Cortisol meets GARP: the effect of stress on economic rationality. Exp. Econ. https://doi.org/10.1007/s10683-019-09624-z.

Choi, S., Kariv, S., Müller, W., Silverman, D., 2014. Who Is (More) rational? Am. Econ. Rev. 104 (6), 1518–1550. https://doi.org/10.1257/aer.104.6.1518.

Chung, H.-K., Tymula, A., Glimcher, P., 2017. The reduction of ventrolateral prefrontal cortex gray matter volume correlates with loss of economic rationality in aging, 1171–17 J. Neurosci.: Off. J. Soc. Neurosci. 37 (49), 12068–12077. https://doi.org/10.1523/JNEUROSCI.1171-17.2017.

Clare, S., Helps, S., Sonuga-Barke, E.J., 2010. The quick delay questionnaire: a measure of delay aversion and discounting in adults. ADHD Atten. Deficit Hyperact. Disord. 2 (1), 43–48.

Cornelisse, S., Van Ast, V., Haushofer, J., Seinstra, M., Joels, M., 2013. Time-Dependent Effect of Hydrocortisone Administration on Intertemporal Choice.

von Dawans, B., Kirschbaum, C., Heinrichs, M., 2011. The Trier Social Stress Test for Groups (TSST-G): a new research tool for controlled simultaneous social stress exposure in a group format. Psychoneuroendocrinology 36 (4), 514–522. https://doi.org/10.1016/j.psyneuen.2010.08.004.

De Kloet, E.R., Joëls, M., Holsboer, F., 2005. Stress and the brain: from adaptation to disease. Nat. Rev. Neurosci. 6 (6), 463–475.

Dias-Ferreira, E., Sousa, J.C., Melo, I., Morgado, P., Mesquita, A.R., Cerqueira, J.J., Costa, R.M., Sousa, N., 2009. Chronic stress causes frontostriatal reorganization and affects decision-making. Science 325 (5940), 621–625. https://doi.org/10.1126/science.1171203.

Feng, C., Wang, H., Lu, N., Chen, T., He, H., Lu, Y., Tu, X.M., 2014. Log-transformation and its implications for data analysis. Shanghai Arch. Psychiatry 26 (2), 105–109.

Frederick, S., 2005. Cognitive reflection and decision making. J. Econ. Perspect. 19 (4), 25–42. https://doi.org/10.1257/089533005775196732.

Friedman, A., Homma, D., Bloem, B., Gibb, L.G., Amemori, K., Hu, D., Delcasso, S., Truong, T.F., Yang, J., Hood, A.S., Mikofalvy, K.A., Beck, D.W., Nguyen, N., Nelson, E.D., Toro Arana, S.E., Vorder Bruegge, R.H., Goosens, K.A., Graybiel, A.M., 2017. Chronic stress alters striosome-circuit dynamics, leading to aberrant decision-making. e28 Cell 171 (5), 1191–1205. https://doi.org/10.1016/j.cell.2017.10.017.

Gathmann, B., Pawlikowski, M., Schöler, T., Brand, M., 2014. Performing a secondary executive task with affective stimuli interferes with decision making under risk conditions (Scopus). Cogn. Process. 15 (2), 113–126. https://doi.org/10.1007/s10339-013-0584-y.

Griefahn, B., Künemund, C., Bröde, P., Mehnert, P., 2001. Zur Validität der deutschen Übersetzung des Morningness-Eveningness-Questionnaires von Horne und Östberg: the validity of a German version of the Morningness-Eveningness-Questionnaire Developed by Horne and Östberg. Somnologie 5 (2), 71–80.

Habhab, S., Sheldon, J.P., Loeb, R.C., 2009. The relationship between stress, dietary restraint, and food preferences in women. Appetite 52 (2), 437–444.

Harbaugh, W.T., Krause, K., Berry, T.R., 2001. GARP for kids: on the development of rational choice behavior. Am. Econ. Rev. 91 (5), 1539–1545. https://doi.org/10.1257/aer.91.5.1539.

Hermans, E.J., van Marle, H.J., Ossewaarde, L., Henckens, M.J., Qin, S., van Kesteren, M.T., Schoots, V.C., Cousijn, H., Rijpkema, M., Oostenveld, R., 2011. Stress-related noradrenergic activity prompts large-scale neural network reconfiguration. Science 334 (6059), 1151–1153.

Hermans, E.J., Henckens, M.J., Joëls, M., Fernández, G., 2014. Dynamic adaptation of large-scale brain networks in response to acute stressors. Trends Neurosci. 37 (6), 304–314.

Het, S., Rohleder, N., Schoofs, D., Kirschbaum, C., Wolf, O.T., 2009. Neuroendocrine and psychometric evaluation of a placebo version of the 'trier social stress test'. Psychoneuroendocrinology 34 (7), 1075–1086. https://doi.org/10.1016/j.psyneuen.2009.02.008.

Houthakker, H.S., 1950. Revealed preference and the utility function. Economica 17 (66), 159–174. https://doi.org/10.2307/2549382.

Jeffreys, H., 1998. The Theory of Probability. OUP, Oxford.

Kimura, K., Izawa, S., Sugaya, N., Ogawa, N., Yamada, K.C., Shirotsuki, K., Mikami, I., Hirata, K., Nagano, Y., Hasegawa, T., 2013. The biological effects of acute psychosocial stress on delay discounting. Psychoneuroendocrinology 38 (10), 2300–2308.

Kirschbaum, C., Pirke, K.-M., Hellhammer, D.H., 1993. The 'trier social stress test' – a tool for investigating psychobiological stress responses in a laboratory setting. Neuropsychobiology 28 (1–2), 76–81. https://doi.org/10.1159/000119004.

Krohne, H.W., Egloff, B., Kohlmann, C.-W., Tausch, A., 1996. Untersuchungen mit einer deutschen Version der 'Positive and Negative Affect Schedule' (PANAS). Diagn. -Gottingen 42, 139–156.

Kudielka, B.M., Kirschbaum, C., 2005. Sex differences in HPA axis responses to stress: a review. Biol. Psychol. 69 (1), 113–132.

Kurtz-David, V., Persitz, D., Webb, R., Levy, D.J., 2019. The neural computation of inconsistent choice behavior. Nat. Commun. 10 (1), 1–14. https://doi.org/10.1038/s41467-019-09343-2.

Landolt, K., Maruff, P., Horan, B., Kingsley, M., Kinsella, G., O'Halloran, P.D., Hale, M.W., Wright, B.J., 2017. Chronic work stress and decreased vagal tone impairs decision making and reaction time in jockeys. Psychoneuroendocrinology 84, 151–158. https://doi.org/10.1016/j.psyneuen.2017.07.238.

Lehrl, S., Triebig, G., Fischer, B., 1995. Multiple choice vocabulary test MWT as a valid and short test to estimate premorbid intelligence. Acta Neurol. Scand. 91 (5), 335–345.

Linden, D.V.D., Keijsers, G.P.J., Eling, P., Schaijk, R.V., 2005. Work stress and attentional difficulties: an initial study on burnout and cognitive failures. Work Stress 19 (1), 23–36. https://doi.org/10.1080/02678370500065275.

Maier, S.U., Makwana, A.B., Hare, T.A., 2015. Acute stress impairs self-control in goal-directed choice by altering multiple functional connections within the brain's decision circuits. Neuron 87 (3), 621–631.

Margittai, Z., Strombach, T., Joëls, M., Schwabe, L., Kalenscher, T., 2015. A friend in need: time-dependent effects of stress on social discounting in men. Horm. Behav. 73, 75–82.

Margittai, Z., Nave, G., Strombach, T., van Wingerden, M., Schwabe, L., Kalenscher, T., 2016. Exogenous cortisol causes a shift from deliberative to intuitive thinking. Psychoneuroendocrinology 64, 131–135. https://doi.org/10.1016/j.psyneuen.2015.11.018.

Margittai, Z., Van Wingerden, M., Schnitzler, A., Joëls, M., Kalenscher, T., 2018. Dissociable roles of glucocorticoid and noradrenergic activation on social discounting. Psychoneuroendocrinology 90, 22–28.

Margittai, Z., Nave, G., Van Wingerden, M., Schnitzler, A., Schwabe, L., Kalenscher, T., 2018. Combined effects of glucocorticoid and noradrenergic activity on loss aversion. Neuropsychopharmacology 43 (2), 334–341. https://doi.org/10.1038/npp.2017.75.

Marin, M.-F., Lord, C., Andrews, J., Juster, R.-P., Sindi, S., Arsenault-Lapierre, G., Fiocco, A.J., Lupien, S.J., 2011. Chronic stress, cognitive functioning and mental health. Neurobiol. Learn. Mem. 96 (4), 583–595. https://doi.org/10.1016/j.nlm.2011.02.016.

Miller, R., Stalder, T., Jarczok, M., Almeida, D.M., Badrick, E., Bartels, M., Boomsma, D.I., Coe, C.L., Dekker, M.C.J., Donzella, B., Fischer, J.E., Gunnar, M.R., Kumari, M., Lederbogen, F., Power, C., Ryff, C.D., Subramanian, S.V., Tiemeier, H., Watamura, S.E., Kirschbaum, C., 2016. The CIRCORT database: reference ranges and seasonal changes in diurnal salivary cortisol derived from a meta-dataset comprised of 15 field studies. Psychoneuroendocrinology 73, 16–23. https://doi.org/10.1016/j.psyneuen.2016.07.201.

Nater, U.M., Rohleder, N., 2009. Salivary alpha-amylase as a non-invasive biomarker for the sympathetic nervous system: current state of research. Psychoneuroendocrinology 34 (4), 486–496.

Nitsch, F.J., Kalenscher, T., 2020. ([Preprint]. PsyArXiv)Keep. a Cool Head. all. What determines Choice consistency? doi: 10.31234/osf.io/etyhx.

Petrowski, K., Paul, S., Albani, C., Brähler, E., 2012. Factor structure and psychometric properties of the trier inventory for chronic stress (TICS) in a representative german sample. BMC Med. Res. Methodol. 12 (1), 42. https://doi.org/10.1186/1471-2288-12-42.

Qin, S., Hermans, E.J., van Marle, H.J.F., Luo, J., Fernández, G., 2009. Acute psychological stress reduces working memory-related activity in the dorsolateral prefrontal cortex. Biol. Psychiatry 66 (1), 25–32. https://doi.org/10.1016/j.biopsych.2009.03.006.

Rammstedt, B., John, O.P., 2007. Measuring personality in one minute or less: a 10-item short version of the big five inventory in English and German. J. Res. Personal. 41 (1), 203–212. https://doi.org/10.1016/j.jrp.2006.02.001.

Rieskamp, J., Busemeyer, J.R., Mellers, B.A., 2006. Extending the bounds of rationality: evidence and theories of preferential choice. J. Econ. Lit. 44 (3), 631–661. https://doi.org/10.1257/jel.44.3.631.

Riis-Vestergaard, M.I., van Ast, V., Cornelisse, S., Joëls, M., Haushofer, J., 2018. The effect of hydrocortisone administration on intertemporal choice. Psychoneuroendocrinology 88, 173–182.

Rohleder, N., Nater, U.M., Wolf, J.M., Ehlert, U., Kirschbaum, C., 2004. Psychosocial stress-induced activation of salivary alpha-amylase: an indicator of sympathetic activity? Ann. N.Y. Acad. Sci. 1032 (1), 258–263. https://doi.org/10.1196/annals.1314.033.

Rouder, J.N., Morey, R.D., Speckman, P.L., Province, J.M., 2012. Default bayes factors for ANOVA designs. J. Math. Psychol. 56 (5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001.

Samuelson, P.A., 1938. A note on the pure theory of consumer's behaviour. Economica 5 (17), 61–71.

Schweda, A., Faber, N.S., Crockett, M.J., Kalenscher, T., 2019. The effects of psychosocial stress on intergroup resource allocation. Sci. Rep. 9 (1), 18620. https://doi.org/10.1038/s41598-019-54954-w.

Starcke, K., Brand, M., 2016. Effects of stress on decisions under uncertainty: a meta-analysis. Psychol. Bull. 142 (9), 909–933. https://doi.org/10.1037/bul0000060.

Stöber, J., 1999. Die soziale-erwünschtheits-skala-17 (SES-17): entwicklung und erste befunde zu reliabilität und validität. The social desirability scale-17 (SDS-17): development and first findings on reliability and validity. Diagnostica 45 (4), 173–177.

Strobel, A., Beauducel, A., Debener, S., Brocke, B., 2002. Is auditory evoked potential augmenting/reducing affected by acute tryptophan depletion? Biol. Psychol. 59, 121–133.

Sugden, R., 1991. Rational choice: a survey of contributions from economics and philosophy (JSTOR). Econ. J. 101 (407), 751–785. https://doi.org/10.2307/2233854.

Tasker, J.G., Di, S., Malcher-Lopes, R., 2006. Rapid glucocorticoid signaling via membrane-associated receptors. Endocrinology 147 (12), 5549–5556.

Varian, H.R., 1982. The nonparametric approach to demand analysis. Econometrica 50, 945–973. https://doi.org/10.2307/1912771.

Varian, H.R., 1991. Goodness-of-fit for Revealed Preference Tests. Department of Economics, University of Michigan Ann Arbor,.

Vinkers, C.H., Zorn, J.V., Cornelisse, S., Koot, S., Houtepen, L.C., Olivier, B., Verster, J.C., Kahn, R.S., Boks, M.P., Kalenscher, T., 2013. Time-dependent changes in altruistic punishment following stress. Psychoneuroendocrinology 38 (9), 1467–1475.

Von Dawans, B., Fischbacher, U., Kirschbaum, C., Fehr, E., Heinrichs, M., 2012. The social dimension of stress reactivity: acute stress increases prosocial behavior in humans. Psychol. Sci. 23 (6), 651–660.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q.F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J.N., Morey, R.D., 2018. Bayesian inference for psychology. Part I: theoretical advantages and practical ramifications. Psychon. Bull. Rev. 25 (1), 35–57. https://doi.org/10.3758/s13423-017-1343-3.

Zellner, D.A., Loaiza, S., Gonzalez, Z., Pita, J., Morales, J., Pecora, D., Wolf, A., 2006. Food selection changes under stress. Physiol. Behav. 87 (4), 789–793.

Study 1b – Trier social stress test and food-choice: Behavioral, self-report & hormonal data

Nitsch, F. J.*, Sellitto, M., & Kalenscher, T. (2021). Trier social stress test and food-choice: Behavioral, self-report & hormonal data. *Data in Brief*, 37, 107245. https://doi.org/10/gmd57p

*Corresponding Author

CRediT Author Statement:

Felix Jan Nitsch: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Original draft preparation, Project Administration

Manuela Sellitto: Data curation, Writing - Review & Editing, Supervision

Tobias Kalenscher: Conceptualization, Resources, Writing - Review & Editing, Supervision, Funding Acquisition

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Trier social stress test and food-choice: Behavioral, self-report & hormonal data

Felix Jan Nitsch*, Manuela Sellitto, Tobias Kalenscher

Comparative Psychology, Institute for Experimental Psychology, Heinrich-Heine-University Düsseldorf, Germany

ARTICLE INFO

Article history:
Received 11 June 2021
Revised 21 June 2021
Accepted 24 June 2021
Available online 25 June 2021

Keywords:
Acute stress
Amylase
Choice Consistency
Cortisol
Chronic Stress
Food-Choice
Rationality

ABSTRACT

A sample of 144 participants underwent the Trier Social Stress Test (TSST), a psychosocial stress manipulation involving a mock interview and a mental arithmetic task, or a matched control procedure. Physiological stress was estimated via a collection of 7 saliva samples over the course of the experiment analysed for cortisol and alpha-amylase, as well as via the mean heart-rate measured before and during the experimental manipulation. Subjective stress was assessed via the Positive and Negative Affect Scale as well as four Visual Analogue Scales at 6 points over the time course of the experiment. Participants solved an incentive-compatible food-choice task before, immediately after and in the aftermath of the experimental manipulation. In each trial of the food-choice task, participants had to choose one out of a set of two to seven snack bundles. Each snack bundle consisted of specific amounts of a sweet or salty snack and a fruit or vegetable. The snacks for both categories were selected to be similarly attractive according to the previously provided online ratings of the participants. The design of the food-choice task allows for the calculation of revealed preference consistency indices. The dataset further contains several self-report questionnaires administered to the participants before the experimental session, including the Trier Inventory of Chronic Stress.

DOI of original article: 10.1016/j.psyneuen.2021.105289

* Corresponding authors.

E-mail address: felix.nitsch@hhu.de (F.J. Nitsch).

Social media: 🐦 (F.J. Nitsch), 🐦 (M. Sellitto), 🐦 (T. Kalenscher)

https://doi.org/10.1016/j.dib.2021.107245

Specifications Table

Subject	Neuropsychology and Physiological Psychology
Specific subject area	Psychoneuroendocrinology, Behavioral Economics, Psychology of Stress
Type of data	Primary data
How data were acquired	Data was gathered using an online survey platform (Unipark), as well as computerized tasks and pen-and-paper questionnaires in a laboratory for measurement of behavior. English translations of all materials are available in the online repository.
Data format	Raw Analysed Figures
Parameters for data collection	The data was obtained from 144 participants in Germany. Participants did not have formal psychological or economic education, were 18–40 years old, non-smokers and did not take medication that could have influenced their corticosteroid levels. Women were not taking oral contraceptives. Similar to previous studies, participants had to refrain from drinking alcohol and sexual activities for 24 h, caffeine four hours and eating/drinking (except water) two hours prior to the beginning of the experiment.
Description of data collection	Participants were recruited via flyers on the university campus, postings in student Whatsapp and Facebook groups and the university job portal (convenience sample).
Data source location	Institution: Heinrich-Heine-University Düsseldorf City/Town/Region: Düsseldorf, Northrhine-Westphalia Country: Germany Latitude and longitude for collected samples/data: 51.233334, 6.783333
Data accessibility	Repository name: Open Science Framework (OSF) Data identification number: DOI 10.17605/OSF.IO/6MVQ7 Direct URL to data: https://osf.io/6mvq7/
Related research article	Nitsch, F. J., Sellitto, M., & Kalenscher, T. (2021). The effects of acute and chronic stress on choice consistency. Psychoneuroendocrinology, 131, 105289. https://doi.org/10.1016/j.psyneuen.2021.105289

Value of the Data

- The data are useful, as both the analysis of salivary cortisol for a relatively large sample as well as the implementation of an incentive-compatible behavioral task are expensive. Further, data sharing in the field of choice consistency/rationality is still relatively uncommon, making the aggregation of evidence challenging [1].
- The data is valuable to researchers interested in the interplay of physiological and subjective stress. It enables exploratory data-analysis regarding individual differences in stress reactivity and mediators of the stress response.
- The unprocessed behavioral data may be used to investigate stress and behavior interactions in economic choice.
- The pre-processed data might be used for evidence aggregation in the field of choice consistency/rationality in the future.

1. Data Description

This OSF directory contains the raw and processed data described, as well as analysis scripts required to computationally reproduce the results and plots reported in the *related research article*. The structure of the directory is

Analysis

- R Studio Project File
- Analysis-scripts (contains all runnable R script files)
- Data (contains raw data)
  - ○ Food choice-data (contains raw data from food-choice task)
- Output (contains all generated output)
  - ○ Data (contains pre-processed data)
  - ○ Plots (contains all plots)

To repeat the analyses of the *related research article*, follow the instructions in the README file.

Materials

- Questionnaires (contains English translations of the administered questionnaire)
- Snack-Pictures (contains all snack pictures used in the food-choice Task)

Here, we will focus on the description of raw and processed data files.

File: root/data/GARP-TSST-mastersheet.xlsx (*raw data*)
Description: The excel file consists of four worksheets. The first sheet ("Master-sheet") contains the raw data of all measures collected during the experimental session except the food-choice task data. The second sheet ("Code-sheet") contains a description of all columns in the Master-sheet. The third sheet ("Saliva-Sample-Encoding") contains excel code matching the IDs of the saliva samples to corresponding participant IDs. The fourth sheet ("Salivettes-Data") contains the raw data for cortisol and amylase measurements that we received from the commercial analysis lab. Typically, users of the data will only import the first sheet into their analysis environment and refer to the Code sheet for further information.

File: root/data/online_data_cleaned.xlsx (*raw data*)
Description: The excel file consists of two worksheets. The first sheet ("Online-Data") contains the raw data of all measures collected during the online survey before the experimental session. The data was cleaned of all identifying information and the survey ID was replaced by the corresponding participant ID. Hence, the file can be readily merged with the Master-sheet by the participant ID column. The second sheet ("Code-sheet") contains a description of all columns in the first sheet. Typically, users of the data will only import the first sheet into their analysis environment and refer to the Code-sheet for further information.

File: root/data/foodchoice-budgetlines.CSV (*raw data*)
Description: The CSV file contains all 22 possible economic parameter combinations which were sampled in the food-choice task (see Experimental Design, Materials & Methods). Each row of the file represents the economic parameters of one possible trial. The first column "m" contains the untransformed budget. The second column "px" contains the price per piece of the sweet/salty snack in untransformed budget units. The third column "py" contains the price per piece of the fruit/vegetable snack in untransformed budget units. The fourth column "px/py" contains the relative price ratio, that is the steepness of the budgetline.

File: root/data/bronars_simulation_data.csv (*raw data*)
The CSV file contains choice consistency data for 10.000 simulated participants, that have been used to determine the power of our food-choice task design to detect choice consistency violations. Simulated participants solved one measurement (11 trials) of the food-choice task (see Experimental Design, Materials & Methods). Simulated choices were uniform-random among the choice sets. Column 1 and 2 ("filename", "VP") are redundant and contain the participant IDs. Column 3 ("Session") indicates the number of measurements (always 1 here). Column 4 ("violation_count") indicates the number of trials involved in a revealed preference inconsistency [see 2]. Column 5 ("AI") contains the critical cost efficiency index [1,3,4]. Column 6 ("mean_RT")

contains the mean reaction times (arbitrary). Typically, users of the data will only use the information in column 4 and 5 (“violation_count”, “AI”).

File: root/data/foodchoice_data/CC-$VPN_$Measurement_data.csv (*raw data*)

$VPN = Participant ID (101-269). $Measurement = Measurement (1-3). Each CSV file contains the raw and untidy data output of the food-choice task. Generally, the files consist of two blocks. The first block consists of columns 1 to 11 (“0”, “1”, “10”, “2”, “3”, “4”, “5”, “6”, “7”, “8”, “9”). It contains the information for the 11 trials of the food-choice task of the corresponding participant and measurement. Row 1 (“MISSES”) indicates how often participants failed to make a response for a given trial within 60 s. Missed trials were repeated at the end of the measurement in random order, until a response was given. Row 2 (“RT”) contains the reaction time in seconds. Row 3 (“choice_set”) contains all available snack bundles in that trial. Row 4 (“m”) contains the untransformed budget. Row 5 column (“px”) contains the price per piece of the sweet/salty snack in untransformed budget units. Row 6 (“py”) contains the price per piece of the fruit/vegetable snack in untransformed budget units. Row 7 (“sbundle”) contains the chosen snack bundle in that trial. Row 8 (“trial_nr”) contains the trial number (1-11) a given trial was first presented. The second block consists of columns 12 to 17 (“Snack_G”, “Snack_UG”, “Testung”, “VP_Code”, “endtime”, “starttime”). For this this block every row contains the same information (redundant). Column “Snack_G” contains the name of the fruit/vegetable snack. Column “Snack_UG” contains the name of the sweet/salty snack. “Testung” contains the measurement. “VP_Code” contains the participant ID. Column “endtime” contains the endtime of the food-choice task. Column “starttime” contains the starttime of the food-choice task. Typically, users will need to clean and pre-process the data before analysis.

File: root/ouput/data/preprocessed-GARP-TSST-data.csv (*pre-processed data*)

The CSV file contains the data of the first sheet of the mastersheet (root/data/GARP-TSST-mastersheet.xlsx) and three additional columns (“CCEI_1”, “CCEI_2”, “CCEI_3”). These columns contain the critical cost efficiency index for each measurement of the food-choice task. These columns were calculated from the raw food-choice data (root/data/foodchoice_data/CC-$VPN_$Measurement_data.csv) using the pre-processing R script (root/analysis_scripts/01_data_preprocessing.R).

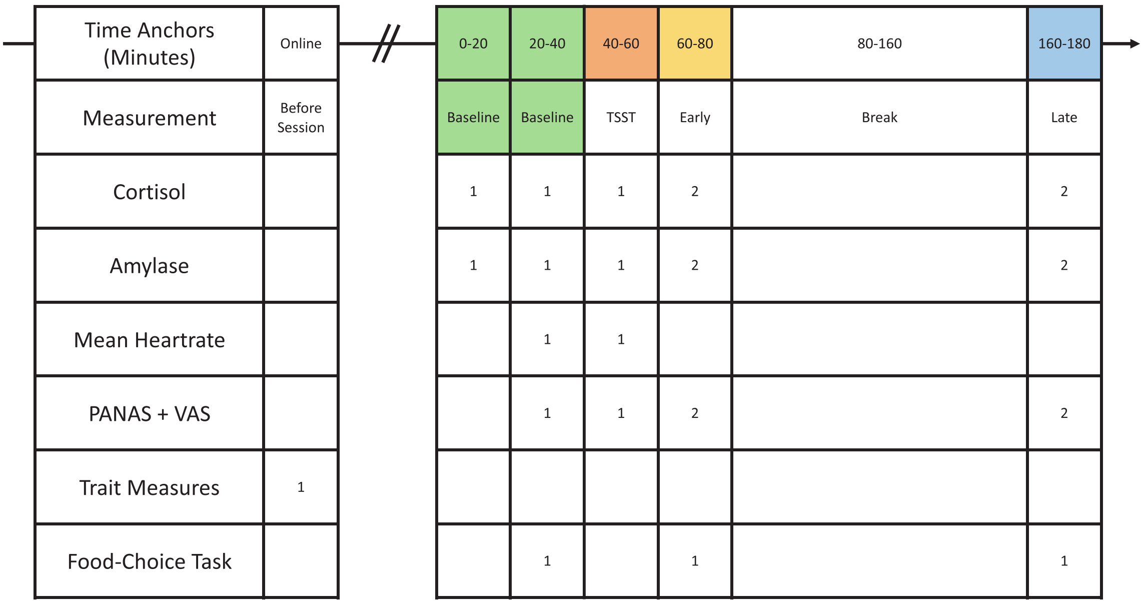
2. Experimental Design, Materials and Methods

The dataset contains data of 144 participants. Participants did not have formal psychological or economic education, were 18–40 years old, non-smokers and did not take medication that could have influenced their corticosteroid levels. Women were not taking oral contraceptives. Similar to previous studies, participants had to refrain from drinking alcohol and sexual activities for 24 h, from caffeine for four hours and from eating/drinking (except water) for two hours prior to the beginning of the experiment.

Fig. 1 provides an exact overview of the experimental timeline for each measure we collected. All experimental sessions took place from 3 p.m. to 6 p.m. to control for circadian variations of hormonal levels. Participants were assigned to the two experimental conditions pseudo-randomly.

Our food-choice task was administered following a 2 × 3 mixed-factorial design with Experimental Group (stress vs. control) as between-subject factor and Measurement (Baseline, Early, Late) as within-subject factor. We deployed a standard food-choice task similar to the one used by Harbaugh et al. [2] and Chung et al. [5]. In each trial, participants chose one out of a set of two to seven snack bundles. Each snack bundle consisted of specific amounts of a sweet or salty snack and a fruit or vegetable (see Fig. 2). The choice set was defined by all integer combinations of sweet or salty snack and fruit or vegetable on the budget-line. The budget-line was given by the following formula:

$$Amount_{Fruit/Vegetable} = -\frac{px}{py}Amount_{Sweet/Salty} + \frac{m}{py}$$



Time Anchors (Minutes)	Online	0-20	20-40	40-60	60-80	80-160	160-180
Measurement	Before Session	Baseline	Baseline	TSST	Early	Break	Late
Cortisol		1	1	1	2		2
Amylase		1	1	1	2		2
Mean Heartrate			1	1			
PANAS + VAS			1	1	2		2
Trait Measures	1						
Food-Choice Task			1		1		1

Fig. 1. Experimental timeline and measurements.
PANAS = Positive and Negative Affect Scale. VAS = Visual Analogue Scales. The figure shows the experimental timeline and measurements. The timeline is sub-split into blocks of twenty minutes, where t = 0 denotes the start of the experimental session in the laboratory. The number within each block (i.e. 1 or 2) denotes the number of measurements taken of the corresponding variable. Trait measures included in the online session were the Trier Inventory of Chronic Stress (TICS), the 10-item version of the Big Five Inventory (BFI-10), the Behavioral Inhibition/Activation Scale (BIS/BAS), the Quick Delay questionnaire (QDQ), the Social Desirability Scale (SDS), the Morningness-Eveningness-Scale (MEQ) and the multiple-choice vocabulary test (MWT-B; German: Multipler Wortschatz Test). Furthermore, participants solved the food rating task in the online session.

At each of the three measurements, participants had to make decisions in 11 trials. The economic parameters ("px", "py", "m") of the 11 trials for each time point were randomly sampled from a collection of 22 possible parameter combinations (see file root/data/foodchoice_budgetlines.csv). The snacks for both categories were similarly attractive according to the previously provided online ratings of the participants (see file root/data/online_data_cleaned.xlsx). The sampling procedure was implemented to reduce interdependency of the answers of each participant for subsequent time points, while keeping the presented bundle size out of satiation range. At the end of the experiment, one trial was randomly selected for each participant, and their choice in that trial was implemented, i.e., participants received their chosen snack bundle. The experimental task was presented via PsychoPy [6]. For 8 participants (out of N = 144 participants), no or incomplete food-choice data were saved due to a technical failure of the experimental hardware.

Our physiological and subjective stress measures were all assessed following a similar mixed-factorial design. However, the exact Measurement regime varied between measures. Self-report trait questionnaires were administered before the experimental session via an online survey (Unipark). An English translation of the administered questionnaires can be found in the online repository.

We collected saliva samples with Salivette devices (Sarstedt, Germany) that consist of a cotton wool swab that participants chewed on lightly for 60 s, allowing the swab to fill with saliva. All samples were frozen and stored at −20 degree Celsius and analysed by a commercial lab (Dresden LabService GmbH) for cortisol and alpha-amylase. Cortisol and alpha-amylase concentrations were determined using a luminescent immunoassay (IBL International, Hamburg). For 16 samples analysis of cortisol and/or alpha-amylase was not possible due to insufficient saliva or contamination with blood.

All analyses were conducted in R [7] in the RStudio IDE [8] using the packages Tidyverse [9], stringi [10], BayesFactor [11] and patchwork [12].

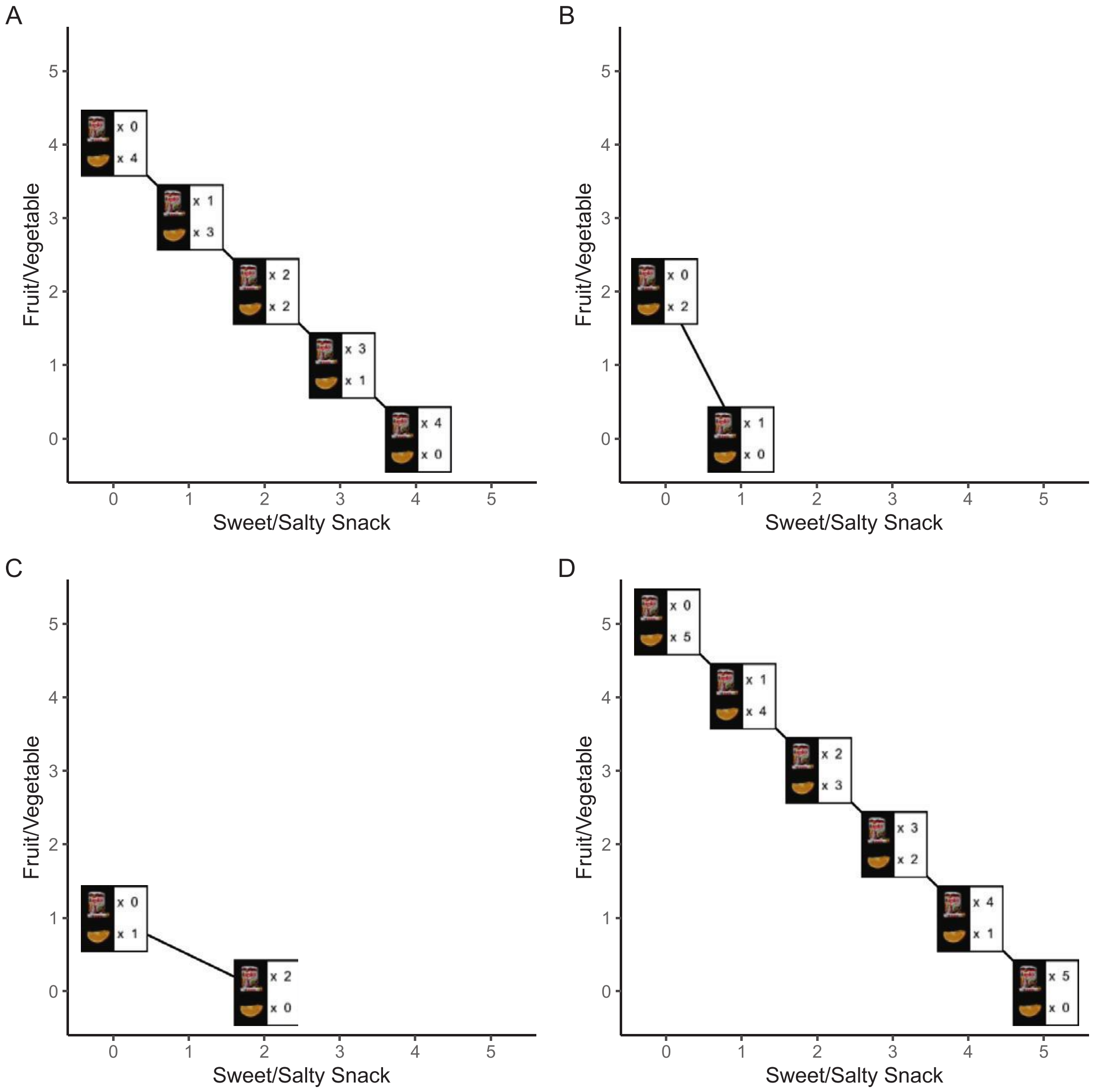


Fig. 2. Four example trials of the Food-Choice Task.

The figure displays the choice sets of 4 example trials of the food-choice task. The economic parameters for these were for A: m = 4, px = 1, py = 1; for B: m = 2, px = 2, py = 1; for C: m = 2, px = 1, py = 2; for D: m = 5, px = 1, py = 1. The available snack bundles are then given by all integer combinations of both snack types along the budgetline. The budgetline is given by the formula:

$$Amount_{Fruit/Vegetable} = -\frac{px}{py} Amount_{Sweet/Salty} + \frac{m}{py}$$

Ethics Statement

All participants gave their informed written consent before participation. The study protocol was approved by the ethical council of the medical faculty of Heinrich-Heine-University Düsseldorf (Study-Nr.: 2020-910). The study was conducted in alignment with the declaration of Helsinki.

CRediT Author Statement

Felix Jan Nitsch: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Original draft preparation, Project Administration; **Manuala Sellitto:** Data curation, Writing - Review & Editing, Supervision; **Tobias Kalenscher:** Conceptualization, Resources, Writing - Review & Editing, Supervision, Funding Acquisition.

Declaration of Competing Interest

The research was supported by a grant from the Deutsche Forschungsgemeinschaft (Bonn, Germany) to Tobias Kalenscher (grant-ID: DFG-KA: DFG-KA 2675/4-3).

The authors declare that they have no known competing financial interests or personal relationships which have or could be perceived to have influenced the work reported in this article.

Acknowledgments

We thank Adam Schweda for helpful comments on the study design as well as assistance with the data collection. We thank Johanna Falk, Anne Schmittner, Jochen Hoog, Alina Keßler and Paul Kramer for assistance with data collection.

Supplementary Materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.dib.2021.107245.

References

[1] F.J. Nitsch, T. Kalenscher, Keeping a cool head at all times. what determines choice consistency? PsyArXiv (2020), doi:10.31234/osf.io/etyhx.

[2] W.T. Harbaugh, K. Krause, T.R. Berry, GARP for kids: on the development of rational choice behavior, Am. Econ. Rev. 91 (2001) 1539–1545, doi:10.1257/aer.91.5.1539.

[3] S.N. Afriat, Efficiency estimation of production functions, Int. Econ. Rev. 13 (1972) 568–598, doi:10.2307/2525845.

[4] H.R. Varian, Goodness of Fit for Revealed Preference Tests, Department of Economics, University of Michigan Ann Arbor, Ann. Arbor., 1993 in.

[5] H.-K. Chung, A. Tymula, P. Glimcher, The reduction of ventrolateral prefrontal cortex gray matter volume correlates with loss of economic rationality in aging, J. Neurosci. 37 (2017) 1171–17, doi:10.1523/JNEUROSCI.1171-17.2017.

[6] J.W. Peirce, PsychoPy—Psychophysics software in Python, J. Neurosci. Methods 162 (2007) 8–13, doi:10.1016/j.jneumeth.2006.11.017.

[7] R Core TeamR: A Language and Environment for Statistical Computing, Vienna, Austria, 2020 https://www.R-project.org/.

[8] RStudio TeamRStudio: Integrated Development for R, RStudio, Inc., Boston, MA, 2018 https://www.rstudio.com/.

[9] H. Wickham, M. Averick, J. Bryan, W. Chang, L. McGowan, R. François, G. Grolemund, A. Hayes, L. Henry, J. Hester, M. Kuhn, T. Pedersen, E. Miller, S. Bache, K. Müller, J. Ooms, D. Robinson, D. Seidel, V. Spinu, K. Takahashi, D. Vaughan, C. Wilke, K. Woo, H. Yutani, Welcome to the Tidyverse, JOSS 4 (2019) 1686, doi:10.21105/joss.01686.

[10] M. Gagolewski, Stringi: Fast and Portable Character String Processing in R, 2021 https://stringi.gagolewski.com/.

[11] R.D. Morey, J.N. Rouder, Bayes Factor: Computation of Bayes Factors for Common Designs, 2018 https://CRAN.R-project.org/package=BayesFactor.

[12] T.L. Pedersen, Patchwork: The Composer of Plots, 2019 https://CRAN.R-project.org/package=patchwork.

Study 2 – Influence of memory processes on choice consistency

Nitsch, F. J., & Kalenscher, T. (2020). *Influence of memory processes on choice consistency.* [Preprint; accepted for publication at *Royal Society Open Science*]. psyarxiv.com/74br5

*Corresponding Author

CRediT Author Statement:

Felix Jan Nitsch: Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Data Curation, Writing - Original draft preparation, Visualization, Project Administration

Tobias Kalenscher: Conceptualization, Resources, Writing - Review & Editing, Supervision, Funding Acquisition

Influence of memory processes on choice-consistency

Felix J. Nitsch and Tobias Kalenscher

Comparative Psychology, Heinrich-Heine-University Düsseldorf, Germany

Author Note

Felix J. Nitsch https://orcid.org/0000-0002-7832-7498

Tobias Kalenscher https://orcid.org/0000-0002-0358-9020

Correspondence concerning this article should be addressed to Felix J. Nitsch, Comparative Psychology, Heinrich-Heine-University Düsseldorf, 40225 Düsseldorf, Germany. Email: felix.nitsch@hhu.de

Abstract

Choice-consistency is considered a hallmark of rational value-based choice. However, because the cognitive apparatus supporting decision-making is imperfect, real decision-makers often show some degree of choice inconsistency. Cognitive models are necessary to complement idealized choice axioms with attention, perception and memory processes. Specifically, compelling theoretical work suggests that the (imperfect) retention of choice-relevant memories might be important for choice-consistency, but this hypothesis has not been tested directly. We used a novel multi-attribute visual choice (MAVC) paradigm to experimentally test the influence of memory retrieval of exemplars on choice-consistency. Our manipulation check confirmed that our retention interval manipulation successfully reduced memory representation strength. Given this, we found strong evidence against our hypothesis that choice-consistency decreases with increasing retention time. However, quality controls indicated that the choice-consistency of our participants was non-discernable from random behavior. In addition, an exploratory analysis showed essentially no test-retest reliability of choice-consistency between two observations. Taken together, this suggests the presence of a floor effect in our data and, thus, low data quality for conclusively evaluating our hypotheses. Further exploration tentatively suggested a high difficulty of discriminating between the choice objects driving this floor effect.

Keywords: Choice-consistency, Memory, Revealed Preference, Cognitive Modeling

Influence of memory processes on choice-consistency

Imagine a stock trader who wants to trade stocks on two different days and plans to invest a starting capital of 600€. On the first day, shares of company A cost 200€ and shares of company B 150€. The stock trader buys 3 shares of company A and 0 shares of company B on the first day. On the second day, the share price of company A sinks to 150€ and the share price of company B rises to 200€. How should the stock trader respond to such a volatile stock market?

A naïve (and inconsistent) stock trader might be tempted to prematurely sell the shares of company A and instead invest into company B. However, this would incur sensitive losses to the trader (de facto 150€, a fourth of the starting capital). More importantly, continuously selling shares cheaper than buying them will inevitably lead to the loss of all capital and being driven out of the market (the so-called *money pump phenomenon*). Such investment behavior might, for example, arise from an inconsistent company value definition. In contrast, a consistent stock trader would base trading decisions on financial analysis, for example considering liquidity, book-to-market value, degree of state-ownership and past performance. This would result in a more robust value definition of company shares than the share price on a given day. Such a stock trader would, ideally, buy stocks at low prices and sell stocks for a profit, using the price volatility advantageously.

Consistent choice can be formalized according to revealed preference theory (Houthakker, 1950; Samuelson, 1938; Varian, 1982). It requires consistent integration of multiple choice attributes (Wallenius et al., 2008) so that it can be rationalized by a monotonous concave utility function (Afriat, 1973). In the example above, utility could be given by the consistent integration of liquidity, book-to-market value, degree of state-ownership and past performance of company shares. Formally, revealed preference theory in its generalized form can be defined as a bound on the structure of the preference relation. Varian (2006) provides a summary of revealed preference theory.

In practice, revealed preference theory is often violated by seemingly inconsistent choice, leading to some researchers proposing sensible relaxations of the choice axioms (Rieskamp et al.,

2006) or the abandonment of choice axioms altogether in favor of a variety of heuristics (Gigerenzer & Selten, 2002).

An important requirement for consistent choice according to revealed preference theory is the stability of preferences and goal structures. The decision maker must have “a definitive structure of wants” (Afriat, 1973). While stability of preferences over prolonged time spans is considered trivial by some (Varian, 2014), others pointed out that preferences may change by endogenous and exogenous cause (Hammond, 1976). Importantly, such dynamic changes of preferences can be the result of natural psychological processes such as attentional shifts, memory encoding and retrieval. Query Theory (Weber et al., 2007) proposes that preferences are not always directly accessible to or completely defined by the decision maker. Instead, relevant experiences with the choice options are retrieved from memory to construct preferences during the decision process: “Preferences, like all knowledge, are subject to the processes and dynamics associated with retrieval from memory” (Johnson et al., 2007). Gabaix & Laibson (2017) propose in a similar notion, that value-based choices are guided by imperfect Bayesian forecasting of future values. These forecasts are derived from prior beliefs and previous experiences. Failure of sufficient retrieval of such memories could result in unstable and incompletely defined preferences and thus, choice inconsistency. Congruently with Query Theory, recent neuropsychological research finds evidence for the relevance of memory-related structures for value-based choices (Wimmer & Shohamy, 2012).

A problem of such choice relevant memory failures is, that they are not directly observable from behavior: We can neither assess which or how well choice relevant memories are retrieved from choice behavior. In a recent preregistered study, Levin et al. (2019) offered a trait heterogeneity based approach to the problem. The authors recruited people who were at least 65 years old to test for the effect of differences in memory abilities (measured by a cognitive assessment battery) on inconsistency in food choice. Participants rated a catalog of food items on a Likert scale. Afterwards, they made repeated pair-wise choices between all possible pairs of food items from the catalog. Memory ability heterogeneity affected divergence of food ratings and actual choices. That is, participants with worse memory ability tended to more frequently choose items with a lower rating over items with a higher rating. However, unexpectedly, memory ability did not influence transitivity of choice itself. It is important to note, that the study by Levin et al. (2019) did not offer any direct measurements of choice-relevant memory retrieval and deploys a non-experimental research design. Therefore, the process of how memory retrieval of goals and preferences affects choice-consistency remains unclear.

In the following sections we will argue that the multi-attribute visual choice paradigm is better suited paradigm to assess the influence of memory retrieval of goals on choice-consistency. Multi-attribute visual choice (MAVC) describes the comparative judgement of visual objects that are characterized by multiple attributes, e.g. orientation, color, shape. Further, we will argue how the revealed preference framework allows a broader evaluation of choice-consistency than traditional accuracy measures of perceptual decisions.

MAVC as a model of value-based choices

In our interpretation, the decision process as postulated by Query Theory (Weber et al., 2007) proposes, at the core, that information about the choice goals is retrieved from memory. Choice options are then compared along all relevant dimensions to the choice goals and the option with maximum integrated goal similarity is chosen. In management science, these goals are also called performance targets (Bordley & Kirkwood, 2004).

For example, when a stock trader decides whether to invest in shares of company A or company B, the trader compares them regarding liquidity, book-to-market value, degree of state-ownership, past performance etc. to a benchmark of what they consider a good investment. The investment option in the choice set which come closest to the benchmark in memory, or so to call choice goals, is chosen (assuming that not investing is no option).

This process is strikingly similar to the decision process proposed by the Generalized Context Model of categorization (Nosofsky, 1986, 2011) in sensory perception, according to which new objects are perceptually categorized based on their similarity to stored exemplars. Exemplars are represented as points in a multi-attribute psychological space. Categorization then is performed by integrating the distance of the new object to the exemplars of a category in memory among all dimensions. Figure 1 shows how important concepts of value-based choice map onto equivalent concepts of multi-attribute visual choice. We propose that the process of comparing choice options to performance goals in value-based choice and to exemplars in MAVC is, psychologically, sufficiently similar to use multi-attribute visual choice as a model for value-based choice.

In MAVC, participants have to choose one out of a set of objects subjectively most similar to a previously learnt exemplar. The choice set objects vary in their similarity to the exemplar regarding multiple attributes. Such choices are comparable to, although not identical with delayed-match-to-sample tasks (Habeck et al., 2004; Steffener et al., 2009, 2012; Zarahn, 2004; Zarahn et al., 2006, 2007).

An important advantage of MAVC tasks over value-based choice tasks is that we can experimentally induce and manipulate exemplar representations in MAVC, whereas goal representations are usually pre-existing, unknown and difficult to manipulate in value-based choice. We can, for example, experimentally manipulate memory representation strength of exemplars through changes of the retention interval between exemplar presentation and choice. These processes are well-studied and several off-the-shelf models for the relationship of memory representation strength and retention interval exist, e.g. exponential and power models (Averell & Heathcote, 2011).

Hypotheses

Based on the predictions of Query Theory (Weber et al., 2007) and neuropsychological evidence on the role of memory for value-based choice (Wimmer & Shohamy, 2012), we expect choice-consistency to be compromised when memory-based goal representations are weak. Correlational evidence partly suggests that this is the case (Levin et al., 2019), however a direct experimental test of the relationship of the strength of memory-based goal representations and choice-consistency is missing and non-trivial to implement.

Based on theoretical considerations (Nosofsky, 1986, 2011; Wallenius et al., 2008) we propose that MAVC can serve as a model for value-based choice. In MAVC, we can experimentally manipulate memory representation strength of exemplars through changes of the retention interval between exemplar presentation and choice. This maps to a manipulation of the strength of memory-based goal representations in our framework (see figure 1). Revealed preference theory (Houthakker, 1950; Samuelson, 1938; Varian, 1982) can be used to analyze MAVC consistency without requiring assumptions about attribute weights or the parametric form of an integration function. Therefore, revealed preference theory can provide a general test of adherence to multi-attribute integration as formulated by the Generalized Context Model of categorization and multi-attribute utility theory. We propose the following hypothesis:

H1**:** As memory representations of exemplars are integral for MAVCs (Nosofsky, 2011), we expect choice-consistency to decrease for longer retention intervals. That is, we expect an inverse relationship of retention interval between learning of the exemplar and choice, and choice-consistency across multiple choices. Hence, we will provide experimental evidence on the role of memory representation strength of goals for choice-consistency.

Previous research on the retention of information shows that forgetting curves are non-linear (Averell & Heathcote, 2011). As we expect choice-consistency to be directly affected by the memory representation, we also expect the relationship of the retention interval and choice-consistency to be non-linear.

H2: We expect choice-consistency to decrease exponentially for longer retention intervals. That is, we expect an exponential model of the relationship of retention interval and choice-consistency to be more strongly supported by the data than a null model (predicting a truncated normal distribution around the mean of the data). The evidence on H2 will help us to quantify the role of memory representation strength of goals for choice-consistency beyond a directional prediction.

H3: In congruence with H2, we expect the exponential decrease of choice-consistency for longer retention intervals to directly replicate in a new data set. This is important, as replicability is a minimal requirement on the meaningfulness of a psychological phenomenon.

Methods

Why is revealed preference theory necessary

Our main dependent variable is consistency in multi-attribute visual choice. We quantified visual choice-consistency with analysis tools borrowed from revealed preference theory. These are preferable over standard indices used in the visual memory and perception literature for conceptual and methodological reasons, as explained in the following.

Value-based choices usually involve trade-offs of different choice attributes. For example, a customer buying snacks might consider both taste and healthiness. While a chocolate bar is arguably tastier, an orange is healthier. A decision, therefore, requires integrating both choice attributes. Whether, taste or health is given more weight is subjective. Concludingly, there is no objectively correct choice. A model of value-based choices should, therefore, include similar attribute trade-offs.

In MAVCs, the choice set stimuli represent a trade-off of similarity to the exemplar regarding multiple attributes. This means that, unlike in traditional memory recognition tasks, such as delayed-match-to-sample tasks, no visual object in the choice set is most similar to the exemplar with regard to all attributes. For example, consider a 3D exemplar cube whose orientation is tilted along the X- and Y-axes (see figure 2). One object in the choice set might be most similar to the exemplar regarding X-orientation while another one is similar regarding Y-orientation. Therefore, there is no objectively correct or dominating choice. This prohibits the use of traditional accuracy measures of perceptual choice that require a normatively correct choice option. In contrast, revealed preference theory allows to test choice-consistency in the context of attribute trade-offs without making unnecessary assumptions about attribute weights or the form of an integration function (Choi et al., 2014).

Sample Characteristics and Exclusion criteria

Participants were recruited from undergraduate psychology students at Heinrich-Heine-University Düsseldorf, Germany on campus and by online adverts. Participants were at least 17 years old, had normal or corrected vision, a good level of German, no neuropsychological or psychiatric diseases and gave informed written consent. The study was approved by the local institutional review board of Heinrich-Heine-University and was conducted in accordance with the declaration of Helsinki. Participants were reimbursed by course credit.

Participants were excluded from the analysis if they do not complete the full experimental session. We did not exclude partial data.

Experimental Setup and Procedure

After participants had given their informed written consent, we assessed age, gender and mother-tongue.

Participants then solved a memory-based visual decision task (see figure 2). In each trial, a 3D exemplar cube was presented for 5 seconds.[1] Each side of the cube was characterized by a unique color in the RGB space[2] from a color scale optimized for color-blind people (Wong, 2011). Each side of the cube was 200px long. The exemplar cube had an orientation of 10, 75, 120, 185, 250 or 315 degrees on the X- and Y-Axis and an orientation of 0 degrees on the Z-Axis. After presentation of the cube a mask of 10 similar cubes (with random X- and Y-orientations) was presented to the participants for a short retention interval. After the retention interval, a choice set of five cubes with

[1] We chose this particular presentation time based on a pilot study (see section Pilot Experiment).

[2] RGB coordinates for each side of the cube. Front: (230, 159, 0). Back: (86, 180, 233). Bottom: (0, 158, 115). Top: (240, 228, 66). Right: (213, 94, 0). Left: (0, 114, 178).

variable X- and Y-orientations was presented, and participants had to select one of the five cubes that had the most similar overall orientation to the exemplar.

The general notion of the task can be compared to that of delayed match-to-sample tasks (Habeck et al., 2004; Steffener et al., 2009, 2012; Zarahn, 2004; Zarahn et al., 2006, 2007) with the difference that there is never a perfect match to the sample. Instead, the choice set stimuli represented a variable trade-off of orientation similarity to the exemplar regarding the X- and Y-axis. For example, a particular stimulus from the choice set might have had a similar X-orientation but a different Y-orientation. Another stimulus might have had a different X-orientation but a similar Y-orientation. Additionally, there could be trials where the choice set stimuli orientations resembled the exemplar orientation more closely and other trials where all choice set stimuli were quite differently oriented from the target stimuli.

The task of the participants was, therefore, to mentally rotate each stimulus of the choice set until it matches the previously shown exemplar and evaluate which of the stimuli required the least mental rotation overall.

Framed in terms of revealed preference theory, each choice trial i was constructed from a budget of $m = 100$ tokens. A pair of prices $p_i = (p_i^{X-ori}, p_i^{X-ori})$ was chosen uniform-randomly from a numeric range of 1 to 3 and 1 to 10. The ranges were assigned to the prices randomly for each trial.

In value-based choice, the price of a good is the cost required to obtain a unit of this good. The budget line then constitutes all combinations of goods affordable spending a fixed budget. Thus, prices and budgets lines are constraints that restrict the possible choice set of combinations of goods out of all available goods. Similarly, the prices and budgets in our multi-attribute choice task constrained the choice set of visual objects out of all possible visual objects characterized by specific attribute values (see figure 3). Given a fixed budget, the prices determined how much ‘similarity’ to the exemplar a participant could ‘purchase‘ along a given orientation axis. The ‘cheaper’ a given dimension, the more similarity to the exemplar on that dimension a participant could afford.

The 5 visual objects were then generated as equidistant-points covering the entire budget line

$$x_i^{Y-ori} = \mathrm{m}/p_i^{Y-ori} - x_i^{X-ori} \times p_i^{X-ori}/p_i^{Y-ori}.$$

Consequently, the choice set always included the extreme objects $x_{i,0} = \left(m/p_i^{X-ori}, 0\right)$ and $x_{i,m} = \left(0, m/p_i^{Y-ori}\right)$. An attribute value of $x_i^{X-ori} = 0$ corresponded to an orientation difference of 30 degrees to the exemplar along the X-axis. With increased values of the X-orientation attribute, the choice object was turned towards the exemplar position along the X-axis. A single unit size amounted to 0.3 degree turn. An attribute value of $x_i^{X-ori} = 100$ corresponded to matching orientation to the exemplar along the X-axis. Likewise, an attribute value of $x_i^{Y-ori} = 0$ corresponded to an orientation difference of 30 degrees to the exemplar along the Y-axis. With increased values of the Y-orientation attribute, the choice object was turned towards the exemplar position along the Y-axis. A single unit size amounted to 0.3 degree turn. An attribute value of $x_i^{y-ori} = 100$ corresponded to matching orientation to the exemplar.

Participants received in-depth instructions about the task (see appendix). Further, they were presented with an animated rotating cube to familiarize with the cube itself.[3] Participants then first solved a practice block of 10 trials with a 1 second retention interval. This practice block served for the participants to familiarize with the design. Choices from the practice block were not included in the analysis. After the practice block, participants were asked to turn to the experimenter in case of questions. Then they solved two consecutive blocks of 20 trials each. For each test block, each participant was assigned a uniform-random retention interval between 0 and 30 seconds (please refer to paragraph “Floor and ceiling effects” below for discussion of the optimal interval length). In total, participants made 20 decisions each for two distinct retention intervals.

After participants had completed the second test block, they solved a similar exemplar reconstruction task as a quality control and manipulation check. The first three screens of each trial

[3] For an impression visit: https://fjnitsch.github.io/files/html/Rotating_Cube.html

were equivalent to the procedure of the main task. Each trial started with the presentation of a fixation cross. Next, participants were presented with an exemplar cube with a certain orientation along the X- and Y-axis for 5 seconds. Then, participants were presented with a mask of 10 randomly oriented cubes (see figure 2) for a certain retention interval. Each participant was assigned a retention interval of either 1, 5, 10 or 30 seconds for the memory reconstruction task. After the retention interval, participants were again presented with a single cube similar to the exemplar. The cube randomly matched the exemplar either regarding the X- or the Y-orientation, while the initial complementary orientation was chosen uniform randomly. Participants then had to turn the cube on the screen to match the exemplar regarding the complementary orientation using the arrow keys on the keyboard. Importantly, it was unknown to the participants whether they would have to reconstruct the X- or Y-orientation both during the presentation time and the retention interval. Participants solved 50 trials of the reconstruction task. Importantly, we did not use the results from the reconstruction task for our main analyses but as a manipulation check.

After completion of the reconstruction task, participants were debriefed about the goals of the study in written form and reimbursed via course credit.

The experimental task was presented with jsPsych (de Leeuw & Motz, 2016). All stimuli were presented on a Lenovo ThinkPad T590 laptop. Subjects were seated 30 cm away from the monitor in a dimly lighted room.

Revealed Preference Theory for MAVC

We measured consistency in MAVC, i.e., the degree of consistency in weighting the two orientation dimensions when comparing the memorized exemplar with the choice set. A participant would act consistent, for example, if they assigned more weight to orientation similarity to the exemplar along one axis when it was expensive, and less weight when it was cheap. Revealed preference theory can be used to quantify the level of inconsistency in weighting the visual attributes in a straight-forward manner.

Let $N \in \mathbb{N}$ be the number of different attributes of a visual object.

Following Nosofsky (2011), let $X = \mathbb{R}_+^N$ be the non-negative, N-dimensional space of visual objects . Let $P = \mathbb{R}_+^N$ be the non-negative, N-dimensional space of prices of attribute similarities to the exemplar. Let $M = \mathbb{R}_+$ be the non-negative, one-dimensional space of budgets. Let $I = i, j \dots \in \mathbb{N}$ denote observations of choice.

Let $x_i \in X$ be the chosen visual object of an observation $i \in I$. Each visual object x_i is a N-dimensional vector of the shape $x_i = (x_i^1, x_i^2, \dots, x_i^N)$, with each scalar component x_i^n representing the similarity of the visual object x_i with regard to attribute n.

Let $p_i \in P$ be the given prices of attribute similarities of an observation $i \in I$. Each prices p_i are a N-dimensional vector of the shape $p_i = (p_i^1, p_i^2, \dots, p_i^N)$, with each scalar component p_i^n representing the price of similarity to the exemplar with regard to attribute n per unit size.

Then the scalar product $x_i p_j$ represents the total price of a visual object x_i at some prices p_j. Let $m_i \in M$ be the given budget of an observation $i \in I$. We assume, that a decision maker spends all her budget so that $x_i p_i = m_i \ \forall i \in I$.

Definition 1 (Direct Revealed Visual Preference). A visual object x_i is directly revealed preferred to another visual object x_j if and only if $x_j p_i \leq m_i$ and $x_i \neq x_j$. Then we denote $x_i R_D x_j$.

Definition 2 (Revealed Visual Preference). A visual object x_i is revealed preferred to another visual object x_j if there exists a transitive preference relation $x_i R_D x_k$, $x_k R_D x_l \dots x_m R_D x_n$, $x_n R_D x_j$ between both bundles. We denote $x_i R x_j$. R is the transitive closure of R_D.

Definition 3 (Strict Direct Revealed Visual Preference). A visual object x_i is strictly directly revealed preferred to another visual object x_j if and only if $x_j p_i < m_i$. Then we denote $x_i P_D x_j$.

Axiom 1 (Generalized Axiom of Revealed Visual Preference). $x_i R \, x_j \rightarrow \neg(x_j P_D x_i) \ \forall i, j \in I$.

Axiom 1 allows us to directly test multi-attribute perceptual choices for consistency. It is a necessary and sufficient condition for the choices to be rationalized by a monotonous concave attribute integration function and, thus, adherence to the Generalized Context Model (GCM) of categorization. If the choice data pass Axiom 1, this means that choices are made as if integrated subjective similarity to the exemplar is a function of objective similarity along each attribute dimension (see figure 3). A simple example of such an integration function could be that subjective similarity is the weighed sum of the similarity along each attribute dimension. As one anonymous reviewer correctly pointed out, mental rotation may not necessarily be performed in an independent, piecewise fashion but possibly also in a holistic mode (Shepard & Metzler, 1971), at least for some participants (Heil & Jansen-Osmann, 2008). We want to emphasize that any concave monotonous similarity function is consistent with revealed preference theory. Hence, an independent (i.e. additive) treatment of the two rotation axes is not required for our model.

However, contrary to Nosofsky (1986), we do not need to make assumptions regarding the parametric form of such an integration function. Conversely, if the data do not pass *Axiom 1* no GCM-style integration function of any monotonous concave specification can rationalize the data.

Preprocessing

For each test block and participant, we calculated the critical cost efficiency index (CCEI; Afriat, 1972, 1973; Varian, 1991). The critical cost efficiency index can be interpreted as how consistently multiple attributes of choice options are integrated into a decision value. The CCEI denotes the “amount by which each budget constraint must be adjusted in order to remove all violations of GARP” (Choi et al., 2007, p. 1927). Computationally, the CCEI presents a relaxation of Axiom 1, so that only $x_i R\ x_j \rightarrow \neg\left(x_j p_j \times \mathrm{CCEI} > x_i p_j\right)\ \forall i, j \in I$ must hold. It ranges from zero to one. A value of one denotes perfect consistency: The attributes are weighed consistently across all choices. The critical cost efficiency index approaches zero as choices become increasingly inconsistent, which means that choice option attributes are weighed inconsistently across different trials. The critical cost efficiency index is the most common indicator of compliance with choice-consistency as defined by revealed preference theory and has been applied in value based choice in various domains (Nitsch & Kalenscher, 2020). Further, we explored the robustness of our results using similar indices such as the money pump index (Echenique et al., 2011), the Houtman-Maks-Index (Heufer & Hjertstrand, 2015) and the minimum cost index (Dean & Martin, 2016). However, since all of these metrics measure slightly different constructs we restrained our preregistered analysis to the critical cost efficiency index.

Analysis Pipeline

Per participant, the data from one test block was randomly selected for testing for an inverse relationship of retention interval and choice-consistency, Bayes factor model comparison and parameter estimation. We call this data *training set*. The other test block was used to replicate our results in a new data set. Therefore, this data was not used for other analyses. We call this data *test set*.

For all analyses, we used a Bayesian framework of inference. Bayesian statistics allows us to express confidence that a parameter is within a certain range, to extend parameter estimation naturally for complicated models, to express evidence for or against hypotheses on a continuous scale and to monitor evidence accumulation (Wagenmakers et al., 2018).

All our analysis were conducted in RStudio (RStudio Team, 2018). We used the following R packages: BayesFactor (Morey & Rouder, 2018), runjags (Denwood, 2016), Tidyverse (Wickham et al., 2019) and patchwork (Pedersen, 2019). Further, we used the JAGS software (Plummer, 2003) for analysis of Bayesian graphical models.

H1: Test for an inverse relationship of retention interval and choice-consistency.

In order to test for an inverse relationship of retention interval and choice-consistency, we calculated Kendall's Tau in the training set. Compared to Pearson's r, it is robust to outliers and violations of normality and expresses dependence in terms of monotonicity instead of linearity (van Doorn et al., 2018). This is important, as we neither expected choice-consistency (index ranging from 0 to 1) or the retention interval (uniformly sampled from an interval of 0 to 30 seconds) to be normally distributed, nor both variables to have a linear relationship. We followed the exact procedure proposed by van Doorn et al. (2018) to test for an inverse relationship of retention interval and choice-consistency using Bayes Factor analysis for Kendall's Tau.

H2: Bayes Factor model comparison of exponential and null model.

In order to gain further insights into the relationship of retention interval and choice-consistency we planned to test which model is supported more strongly by the data of the training set (but see section Interpretative Plan and results for H1 why we did not proceed to test this hypothesis). For this, we planned to use Bayes Factor model comparison via the product space method (Lodewyckx et al., 2011). We planned to test two candidate models against each other, which are specified in the following sections. We assumed both models to had equal prior probabilities.

$p_{M1} = p_{M2} = 0.5$

The first candidate is inspired by forgetting models of item recall (Averell & Heathcote, 2011), the second model is a null model assuming no effect of the retention interval on choice-consistency. They give rise to observed participant choice-consistency, given a retention interval.

Both candidate models are of the general form:

$$CCEI_t \sim Normal(\mu_t, \sigma)$$

$with\ CCEI_t, \sigma \in [0,1].$

$CCEI$ denotes the critical cost efficiency index of a participant for one test block, t denotes the assigned retention interval for that block (ranging from 0 to 30 seconds). σ accounts for random noise in the data. We assume all parameter values for σ to be equally likely a priori.

$$\sigma \sim Beta(1,1)$$

μ_t denotes the expected choice-consistency given a retention interval and is specific to the model candidates.

Exponential Model. The first candidate model assumes that choice-consistency decreases exponentially with retention time. This means that the decreasing rate of consistency is constant over retention time. Following Averell & Heathcote (2011), the function can be formalized in the following way:

$$\mu_t = a + (1 - a) \times b \times e^{-\alpha \times t}$$

The parameter $a \in [0,1]$ determines an asymptotical minimum level of choice-consistency after an infinite retention interval. The parameter $b \in [0,1]$ determines choice-consistency at $t = 0$, which allows for imperfect choice-consistency unconditional on time-dependent processes when $b < 1$. The parameter $\alpha \in [0,1]$ determines the retention time-constant decreasing rate of consistency. We assume that all parameter values are equally likely a priori.

Figure 4 displays a graphical representation of the model including prior specifications for all parameters.

Null Model. The second candidate model assumes that choice-consistency does not decrease as a function of the retention interval. The expected value of the choice-consistency distribution is, therefore, a constant.

$$\mu_t = c$$

The parameter $c \in [0,1]$ determines the expected value of the choice-consistency. We assume that all parameter values for c are equally likely a priori.

$$c \sim Beta(1,1)$$

H3: Replication for the test set

In order to test whether the relative advantage in support by the data for the exponential model in comparison to the null model replicates to a new data set, we planned to obtain the replication Bayes factor using the held out test set using the method described by Ly et al. (Ly et al., 2019; but see section Interpretative Plan and results for H1 why we did not proceed to test this hypothesis). The replication Bayes factor is given by Bayes Factor for the coerced data set divided by the Bayes factor for the training set (obtained for H2).

$$BF_{10}(d_{test}|d_{train}) = \frac{BF_{10}(d_{test}, d_{train})}{BF_{10}(d_{train})}$$

This evidence updating method does not require approximations and is especially useful for complex models as in our application case.

Interpretative Plan

We followed the usual framework (Jeffreys, 1998) for interpreting Bayes Factors, which means that we considered a Bayes factor of $BF \geq 10$ as strong evidence for a hypothesis. Table 1 summarizes the interpretative plan for all hypotheses.

We collected further data until we reach a conclusive result for all hypotheses.

H1: Should we find strong support for an inverse relationship of retention interval and choice-consistency, we would conclude that choice-consistency in MAVC depends on the memory representation strength of exemplars. Should we find strong evidence against an inverse relationship of retention interval and choice-consistency, this would question the role of memory representation of exemplars in MAVC. It could be concluded, that choice-consistency is robust to indefinite goal representations. In this case, we would not proceed to test H2 and H3.

H2: Should we find strong evidence, that the exponential model of the relationship of retention interval and choice-consistency is supported more strongly by the data than the null model, we would interpret this as preliminary evidence for the validity of the exponential model. However, a definitive interpretation would require generalizability of the results for the test set. Furthermore, our statistical tests would only collect relative evidence for one model over another. It would still be possible, that the true model is outside our model space. Therefore, careful inspection of the visualizations of the model predictions would be required (see figures 5 and 6). Should we find strong evidence in support of the null model, this would question the validity of an exponential model specifically, given positive evidence for H1. Again, a definitive interpretation would require generalizability of the results for the test set.

H3: Should we find strong evidence that the relative advantage in support by the data for the exponential model in comparison to the null model replicates to a new data set, we would interpret this as further evidence for the validity of the exponential model. Should the replication Bayes factor favor the null model, this would question the validity of an exponential model specifically, given positive evidence for H1. Again, our statistical tests would only collect relative evidence for one model over another. Careful inspection of the visualizations of the model predictions would be required (see figures 5 and 6).

Should we find conflicting evidence for H2 and H3, we would use the Bayes factor for the complete dataset ($BF_{10}(d_{test}, d_{train})$ to guide our interpretation. The Bayesian model comparison using the complete dataset quantifies the evidence for or against each model in light of all data. We would use the same interpretation framework as before, which means that we consider a Bayes factor of $BF \geq 10$ as conclusive evidence.

Data Collection Plan / Power Analysis

Inferential power

Our data collection plan is based on a Bayesian stopping rule: We collected data until we reached a Bayes factor of $BF \geq 10 \vee BF \leq 0.1$ or a maximum feasible sample size of $N = 500$.

Sensitivity of choice-consistency test

In order to make meaningful statements about the influence of memory processes it is not only necessary to experimentally manipulate these memory processes with a sufficient effect size but also to measure choice-consistency with sufficiently sensitive measure. The sensitivity of our behavioral task to detect violations of choice-consistency can be approximated using a simulation study (Bronars, 1987). We simulated a dataset of 1.000 virtual participants that made uniform random choices from 20 choice sets constructed as specified for our experiment (see Procedure). Results showed that 99% of the virtual participants violated choice-consistency at least once with a median CCEI of 0.389 (see figure 7).

Specification of Reality Checks

First, to ensure that our retention interval manipulation is effective, we tried to replicate the effect of the retention interval on absolute reconstruction error of exemplars from memory that we found in our pilot experiment (see pilot experiment) in our control task. Specifically, we wanted to find strong evidence (Bayes factor of at least BF ≥ 10) favoring a one-way ANOVA style model including the 4-step retention interval factor over a null model. Inference was based on the replication Bayes factor fully utilizing the evidence from our pilot experiment with $BF_{10}(d_{orig}) = 1000$ (Ly et al., 2019).

We used the JAGS software (Plummer, 2003) to analyze our Bayesian graphical models. To assess convergence we used trace plots of the Markov-Chain-Monte-Carlo simulations and smoothed density plots of the parameter estimates (Kruschke, 2014).

Following Blaha (2019), we think that visualization is an important reality check to see, whether the data looks like we expected and hypothesized. Therefore, we planned to create two main plots for visual qualitative checks of the data and models.

First, we created a scatter plot of retention interval and choice-consistency in the training set together with histograms of the marginal distributions. This allowed for a visual inspection of the relationship of both variables as well as the marginal distributions. We did not want to see choice-consistency increase as a function of retention interval, as such pattern is not covered by our model space. The marginal distribution of the retention interval should, trivially, be uniform (as generated by the experimental task). The marginal distribution of the choice-consistency should, ideally, be a right-tailed Gaussian, meaning a tail for large consistency values. Figure 5 shows such a plot for data simulated from the exponential model.

Second, we planned to create a plot that is overlaying scatter plots of retention interval and choice-consistency in the training set with the posterior predictive distributions of the exponential model and the null model (but see section Interpretative Plan and results for H1 why we did not proceed with our computational modelling). This would allow us to visually inspect how well the models can explain the data and further, if there are any important qualitative differences between predictions and data. This is an important step to inform future modelling efforts and to identify systematic short-comings of a model. Further, we would create the same plots for retention interval and choice-consistency in the test set overlaid with the out-of-sample posterior predictions of both models to qualitatively evaluate the generalizability of the models (but see section Interpretative

Plan and results for H1 why we did not proceed with our computational modelling). Figure 6 shows such plots for data simulated from the exponential model.

Floor and ceiling effects

Should we find that choice-consistency is either near perfect or at very low levels across all retention intervals, this would indicate ceiling or floor effects respectively. While this is theoretically possible (e.g. in case of an ineffective retention interval manipulation), we reduced the likelihood of finding such a pattern by using a continuous manipulation of the retention interval instead of a factorial design. Therefore, our design covered a wide range of retention intervals (interval of 30 seconds) instead of 2 to 3 retention intervals, a factorial design would cover. Still, it was not possible to entirely rule out the possibility of an ineffective retention interval manipulation on theoretical grounds only. Therefore, we conducted a pilot experiment to demonstrate the effectiveness of our manipulation using the control task from our main experiment (see below).

Further limitations

As one anonymous reviewer pointed out, our MAVC paradigm does not include a no-choice option. Intuitively, for some trials it would be difficult for participants to make a similarity judgement. However, we decided not to include a no-choice option in our paradigm as one core assumption of revealed preference theory is that there is a well-defined preference structure (Afriat, 1973) and this also holds for difficult decisions. Therefore, asking participants to make a choice for difficult decisions is part of a rigorous test of revealed preference choice-consistency. Still, practically, this could have introduced additional noise into the decision behavior of participants. While the current registered report cannot entirely address this aspect of insufficiently defined preferences, future research should provide both theoretical and empirical accounts on the role of non-decisions for choice-consistency.

Pilot Experiment

We conducted a pilot experiment to validate the effectiveness of our retention interval manipulation. Specifically, we wanted to demonstrate that our retention interval manipulation is sufficient to blur the memory representation of the exemplar. Furthermore, we explored the influence of presentation time of the exemplar on the memory representation strength.

Methods

Procedure. The first three screens of each trial were equivalent to the procedure of the experiment for the here described registered report (see figure 2). Each trial started with the presentation of a fixation cross. Participants were then presented with an exemplar cube with a certain orientation along the X- and Y-axis (see procedure of registered report) for either 1, 5, 10 or 30 seconds. Then, participants were presented with a mask of 10 randomly oriented cubes (see figure 2) for a certain retention interval. Importantly, the retention interval in the pilot experiment was not fixed per participant. The retention interval lasted either 1, 5, 10 or 30 seconds. After the retention interval, participants were again presented with a single cube similar to the exemplar. The cube randomly matched the exemplar either regarding the X- or the Y-orientation, while the initial complementary orientation was chosen uniform randomly. Participants then had to turn the cube on the screen to match the exemplar regarding the complementary orientation using the arrow keys. Importantly, it was unknown to the participants whether they had to reconstruct the X- or Y-orientation both during the presentation time and the retention interval. Participants solved 10 for each factorial combination of the presentation times and retention intervals in random order.

Sample characteristics and exclusion criteria. We included a total of 25 participants (21 women, 4 men; age: $M = 24, Range = 18 - 39$) for our pilot experiment. The sample size was not determined a priori. Instead we used a Bayesian stopping rule, recruiting further participants until we reached a Bayes factor of at least BF ≥ 10 for our hypothesis test. Importantly, the sample size is smaller than the minimal sample size we plan to recruit for the here described registered report. Participants were recruited from the same population we target for the here described registered report.

However, the study was conducted as an online experiment due to the ongoing COVID-19 crisis. It is intuitive, that participants might be less attentive during online-experiments than during lab-based experiments due to the uncontrolled in which participants solve the task. Therefore, we assessed reaction times besides task performance and excluded single trials with reaction times deviating more than 3 standard deviations from the grand mean. Note, that this threshold amounted to about 30 seconds for a single trial. Hence, we are confident to not have excluded any meaningful data while considerably reducing measurement noise.

The study was approved by the local ethics board of Heinrich-Heine-University and was conducted in accordance with the declaration of Helsinki. Participants were reimbursed by course credit.

Statistical analysis. We operationalized the memory representation strength of the exemplar as the absolute error with which its orientation could be reconstructed by the participants. We considered an orientation of 0 degrees and 360 degrees as equivalent. For example, if the orientation of the exemplar cube on the axis of interest is 90 degrees and the orientation of the reconstructed cube on that axis is 360, the absolute error is 90 and not 270. The absolute error can therefore range between 0 degrees and 180 degrees. The exact formula is given by

$$Absolute\ Error = \left| \left| Ori_{Exemplar} - Ori_{Reconstructed} \right| - 180 \right|.$$

We calculated a repeated measures Bayesian ANOVA in the 'BayesFactor' R package using the non-informative default priors (Rouder et al., 2012). We considered a Bayes factor equal or larger than 10 regarding the main effect of the retention interval to be conclusive for or against our hypothesis. To verify the direction of the effect we considered the trend of the means of each factor level. Further, we exploratively inspected the evidence for or against a main effect of the presentation time and a possible interaction effect of both factors.

Results

We found that an ANOVA-style model including both main effects and a random subject intercept but no interaction term to be the most likely model given the data. Specifically, this model was $BF_{10} = 895082$ $(\pm 0.94\%)$ times more likely than the null model (including only the random subject intercept) given the data (see figure 8).

Retention interval. To quantify the evidence for an effect of the retention interval factor we compared the evidence of the most likely model with the evidence for a model including only the main effect of the presentation time and the random subject intercept. We found that there was $BF_{10} = 1000\ \pm 1.13\%$ times more evidence for the inclusion of the retention interval factor. We interpret this as definitive evidence. Inspection of the trend of means reveals that there is a positive relationship of retention interval and absolute error of reconstruction (see figure 8).

Presentation time. To quantify the evidence for an effect of the presentation time factor we compared the evidence for the most likely model with the evidence for a model including only the main effect of the retention interval and the random subject intercept. We found that there was $BF_{10} = 1005\ \pm (1.15\%)$ times more evidence for the inclusion of the presentation time factor. We interpret this as definitive evidence. Inspection of the trend of means reveals that there is a negative relationship of presentation time and absolute error of reconstruction (see figure 8). Note, however, that is was an explorative analysis.

Interaction Retention interval x presentation time. To quantify the evidence against an interaction effect of both factors we compared the evidence of the most likely model with the evidence for a model including both main effects, the random subject intercept and an interaction term. We found that there was $BF_{01} = 44446\ (\pm 2.11\%)$ times more evidence for the exclusion of the interaction term. We interpret this as definitive evidence. Note, however, that is was an explorative analysis.

Discussion

We conducted a pilot experiment to validate the effectiveness of our retention interval manipulation. We showed that the precision of the orientation reconstruction of an exemplar from memory decreases with retention time over an interval of 30 seconds. Therefore, we are confident that the planned retention interval manipulation of the here described registered report is effective to weaken the memory representation strength of an exemplar. Further, we explored the influence of different presentation times on orientation reconstruction precision. We found that precision increases with presentation time. In the context of our registered report, it is important that the memory representation strength freely varies among different retention intervals for a given presentation time. Descriptively, the variance of the absolute error in reconstruction is highest for a presentation time of 1 second. However, also the mean absolute error is highest for a presentation time of 1 second. A presentation time of 5 seconds represents a compromise with the second highest variance and second highest mean of the absolute error in reconstruction.

Results

Sample Characteristics

For our main experiment, we included 77 participants (56 women, 21 men; age: $M = 22, Range = 18 - 40$, education: 72 completed high school, 5 completed a university degree) according to our inclusion criteria until reaching the preregistered stopping rule of our analysis plan.

Preregistered analyses

As outlined above, our statistical analyses use a Bayesian framework of inference, specifically the Bayes Factor approach to model comparison. Bayes factors express the *relative* degree of evidence for one model over another, that is the ratio of probabilities of observing the data under each model (Makowski et al., 2019).

Reality check: Effect of retention interval on memory representation

To quantify the evidence for an effect of the retention interval factor on memory representation strength, we compared a one-way ANOVA style model including the 4-step retention interval factor to a null model. We found conclusive evidence for the retention interval model for the new and the full dataset (including our pilot data), as well as for the successful replication ($BF_{10}(d_{new}) = 32.894\ \pm 0.01\%$, $BF_{10}(d_{full}) = 46717770\ \pm 0.01\%$, $BF_{Replication} = 44623.3$; see figure 9). Hence, we can conclude that our retention interval manipulation was effective.

H1: Test for an inverse relationship of retention interval and choice-consistency

Next, we tested in the training set whether choice-consistency as operationalized by the CCEI decreased with an increasing retention interval. Results showed conclusive evidence against our hypothesis ($BF_{10} = 0.047$; see figure 10). The result held for other specifications of choice-consistency, namely the Houtman-Maks-Index, and approximations of the Money Pump Index and Minimum Cost Index (all $BF_{10} < 0.1$).

H2 & H3: Bayes Factor model comparison of exponential and null model

Following our interpretation plan (see table 1), we did not proceed to test H2 and H3, given our negative result for H1.

Floor and ceiling effects

As apparent in figure 10, the CCEI of our participants in the training data was overall surprisingly low ($Median = 0.299, SD = 0.229$). In order to control for a potential floor effect, we used a Bayesian Mann-Whitney-U test to control whether our participants were more consistent than an equal-sized subsample of our random simulated data (see figure 11, panel A). Results showed strong evidence against this, indicating a potential floor effect in our data ($BF_{10} = 0.094$).

Exploratory analysis

Reliability Analysis

In an attempt to further understand the quality of our data beyond our preregistered quality controls, we conducted a descriptive test-retest reliability analysis of the CCEI for the training and test set. As we reported recently elsewhere (Nitsch et al., 2021), there are concerns regarding the measurement reliability of the CCEI, which is especially problematic for correlational designs such as the one of the current study (Hedge et al., 2018). Specifically, tasks designed to show robust between-group effects and, thus, low between-subject variability in the outcome measure are at risk for showing low test-retest reliability. Another risk factor specific to the CCEI is that the measure is dependent only on the magnitude of the most severe violation (see Preprocessing) and, thus, vulnerable to outliers. Our results indicated essentially no reliability of the CCEI between training and test set in the current study ($r = -0.033$). This was not driven by the difference in retention intervals of both measurements, or by low between-subject variability of the measure (see figure 11, panel A and B). A similar result showed for the money pump index ($r = 0.023$). However, interestingly, the Houtman-Maks-Index and the Minimum-Cost-Index showed a much higher (albeit still poor) test-rest reliability (HMI: $r = 0.303$, MCI: $r = 0.353$; see figure 11, panels C-E), which might be attributed to less vulnerability to outliers.

Task difficulty

Given the results reported above and oral feedback from our participants, we formed the post-hoc hypothesis that the generally low choice-consistency might be driven by a too high difficulty of discriminating the different X- and Y-orientations of the choice objects. To further explore this notion, we compared the mean absolute error in the reconstruction task, as an upper limit to the discriminatory performance, to the mean increment difference of orientation along the X- or Y-axis in the choice set, using bootstrapping. Results showed, that the mean increment difference was generally lower than mean reconstruction error, tentatively suggesting a high difficulty of discriminating between the choice objects (see figure 12).

Discussion

In this registered report, we set out to experimentally test the influence of memory retrieval of exemplars on choice-consistency in a novel visual choice paradigm. After a short retention interval, participants had to select one out of a choice set of five three-dimensional cubes that has the subjectively most similar orientation along the X- and Y-axis to the exemplar. The choice set stimuli represented a variable trade-off of similarity to the exemplar regarding the two attributes X- and Y-orientation. We manipulated memory retrieval by varying the duration of the retention interval between exemplar presentation and choice.

Using a reconstruction task as a manipulation check of our retention interval intervention, we could show and replicate the pattern of decreasing memory accuracy with increasing retention time in a pilot experiment and in our preregistered study, which confirmed the effectiveness and reliability of our manipulation.

Given this, we found strong evidence against our first hypothesis that choice-consistency, as operationalized by the CCEI, decreases with increasing retention time. Further, this result held for robustness checks using three similar choice-consistency indices. Given our preregistered interpretation plan we did not proceed to test our more specific, model-based hypotheses.

Limitations

However, our preregistered quality controls revealed an overall surprisingly low choice-consistency of our participants even for short retention intervals that proved to be non-discernable from that of simulated random behavior. In addition, an exploratory analysis showed essentially no test-retest reliability of the critical cost efficiency index between the training and the test set. This was not driven by retention time differences between the two measurements. Taken together, this suggests the presence of a floor effect in our data and, thus, low data quality for conclusively evaluating our hypotheses.

Generally, the lack and low reliability of choice-consistency indicates that our participants did not consistently integrate deviations in the X- and Y-dimensions of the choice set stimuli to the exemplar, meaning there was no well-behaved integration function, and, this was also the case for short retention times. As the performance in the reconstruction task was generally good (perfect reconstruction in about 42% of trials), it is unlikely that the low consistency level was driven by too long retention intervals.

Another explanation for the overall low choice-consistency could be that the discrimination between the different X- and Y-orientations of the choice objects was too difficult. This was also, anecdotally, suggested in oral feedback of our participants during the data collection. To further explore this notion, we compared the mean absolute error in the reconstruction task, as an upper limit to the discriminatory performance, to the mean increment difference of orientation along the X- or Y-axis in the choice set, recovered from the task parameters, using bootstrapping. Results showed that the mean increment difference was generally lower than mean reconstruction error, tentatively suggesting a high difficulty of discriminating between the choice objects. As all choice-consistency indices quantify performance only relative to the increment orientation differences of choice objects, numerically low choice-consistency levels can correspond to only small inconsistencies in degree orientation.

Future Research

Future studies investigating visual choice-consistency should, therefore, establish a sufficient level of choice-consistency at baseline. This could be achieved, for example, by pilot testing to adjust, in a group-wise fashion, the increment difference of orientation along the X- or Y-axis in the choice, or on an individual-level by using an adaptive staircase procedure.

Another important consideration for the design of future studies is the low reliability of choice-consistency in the present study, but also, generally, in other task domains (Nitsch et al., 2021). While our correlational design had important benefits for covering a sufficient retention time span and providing rich data for parametric model fitting, factorial designs are more robust to finding effects in low reliability behavioral tasks (Hedge et al., 2018).

Conclusion

In this registered report, we set out to experimentally test the influence of memory retrieval of exemplars on choice-consistency in a novel visual choice paradigm. Due to unforeseen methodological pitfalls, our data is inconclusive to the preregistered hypotheses. However, our preregistered quality controls and additional exploratory analyses offer important insights for the design of future studies.

Acknowledgements

We thank Alina Keßler, Paul Kramer and Marie Ludes for their support on the pilot data collection. We thank Quentin Gronau and three anonymous reviewers for their helpful criticism which led to a greatly improved version of stage 1 manuscript. Further, we thank Ana Hernandez for her support on the data collection of the main experiment.

Ethical statement

All participants gave informed written consent. The study was approved by the local institutional review board of Heinrich-Heine-University and was conducted in accordance with the declaration of Helsinki. Participants were reimbursed by course credit.

Funding statement

The research was supported by a grant from the Deutsche Forschungsgemeinschaft (Bonn, Germany) to Tobias Kalenscher (grant-ID: DFG-KA 2675/4-3).

Data Availability

Raw and processed data, the approved stage 1 manuscript, as well as analysis code for the stage 1 and stage 2 manuscript results and figures are available online: https://osf.io/2vx36/.

Competing Interests

We have no competing interests.

CRediT author statement

Felix Jan Nitsch: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization, Project administration. **Tobias Kalenscher:** Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

References

Afriat, S. N. (1972). Efficiency Estimation of Production Functions. *International Economic Review*, *13*(3), 568–598. JSTOR. https://doi.org/10/bkn49z

Afriat, S. N. (1973). On a system of inequalities in demand analysis: An extension of the classical method. *International Economic Review*, 460–472. https://doi.org/10/fdr8kn

Averell, L., & Heathcote, A. (2011). The form of the forgetting curve and the fate of memories. *Journal of Mathematical Psychology*, *55*(1), 25–35. https://doi.org/10/c4fp56

Blaha, L. M. (2019). We Have Not Looked at Our Results Until We Have Displayed Them Effectively: A Comment on Robust Modeling in Cognitive Science. *Computational Brain & Behavior*, *2*(3–4), 247–250. https://doi.org/10/ggdv5w

Bordley, R. F., & Kirkwood, C. W. (2004). Multiattribute Preference Analysis with Performance Targets. *Operations Research*, *52*(6), 823–835. https://doi.org/10/bz6xqk

Bronars, S. G. (1987). The power of nonparametric tests of preference maximization. *Econometrica: Journal of the Econometric Society*, 693–698. https://doi.org/10/drb9tp

Choi, S., Fisman, R., Gale, D., & Kariv, S. (2007). Consistency and heterogeneity of individual behavior under uncertainty. *American Economic Review*, *97*(5), 1921–1938. https://doi.org/10/c3665n

Choi, S., Kariv, S., Müller, W., & Silverman, D. (2014). Who Is (More) Rational? *American Economic Review*, *104*(6), 1518–1550. https://doi.org/10/76w

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*, *48*(1), 1–12. https://doi.org/10/f8hb63

Dean, M., & Martin, D. (2016). Measuring rationality with the minimum cost of revealed preference violations. *Review of Economics and Statistics*, *98*(3), 524–534. https://doi.org/10/gdz4xw

Denwood, M. J. (2016). runjags: An R Package Providing Interface Utilities, Model Templates, Parallel Computing Methods and Additional Distributions for MCMC Models in JAGS. *Journal of Statistical Software*, *71*(9). https://doi.org/10/gf2rp9

Echenique, F., Lee, S., & Shum, M. (2011). The money pump as a measure of revealed preference violations. *Journal of Political Economy*, *119*(6), 1201–1223. https://doi.org/10/f3wmsn

Gabaix, X., & Laibson, D. (2017). *Myopia and discounting*. National bureau of economic research.

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press. https://doi.org/10.7551/mitpress/1654.001.0001

Habeck, C., Rakitin, B. C., Moeller, J., Scarmeas, N., Zarahn, E., Brown, T., & Stern, Y. (2004). An event-related fMRI study of the neurobehavioral impact of sleep deprivation on performance of a delayed-match-to-sample task. *Cognitive Brain Research*, *18*(3), 306–321. https://doi.org/10/bd69xn

Hammond, P. J. (1976). Changing Tastes and Coherent Dynamic Choice. *The Review of Economic Studies*, *43*(1), 159–173. JSTOR. https://doi.org/10/d5sfq6

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10/gddfm4

Heil, M., & Jansen-Osmann, P. (2008). Sex Differences in Mental Rotation with Polygons of Different Complexity: Do Men Utilize Holistic Processes whereas Women Prefer Piecemeal Ones? *Quarterly Journal of Experimental Psychology*, *61*(5), 683–689. https://doi.org/10/fnrqjr

Heufer, J., & Hjertstrand, P. (2015). Consistent subsets: Computationally feasible methods to compute the Houtman–Maks-index. *Economics Letters*, *128*, 87–89. https://doi.org/10/f68m4d

Houthakker, H. S. (1950). Revealed preference and the utility function. *Economica*, *17*(66), 159–174. https://doi.org/10/fhq8cj

Jeffreys, H. (1998). *The theory of probability*. OUP Oxford.

Johnson, E. J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: A query theory of value construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 461–474. https://doi.org/10/dcgsrs

Kruschke, J. (2014). MCMC representativeness. In *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (pp. 178–181). Academic Press.

Levin, F., Fiedler, S., & Weber, B. (2019). The influence of episodic memory decline on value-based choice. *Aging, Neuropsychology, and Cognition*, *26*(4), 599–620. https://doi.org/10/gk9pvr

Lodewyckx, T., Kim, W., Lee, M. D., Tuerlinckx, F., Kuppens, P., & Wagenmakers, E.-J. (2011). A tutorial on Bayes factor estimation with the product space method. *Journal of Mathematical Psychology*, *55*(5), 331–347. https://doi.org/10/bc87rf

Ly, A., Etz, A., Marsman, M., & Wagenmakers, E.-J. (2019). Replication Bayes factors from evidence updating. *Behavior Research Methods*, *51*(6), 2498–2508. https://doi.org/10/gg7gzx

Makowski, D., Ben-Shachar, M. S., Chen, S., & Lüdecke, D. (2019). Indices of effect existence and significance in the Bayesian framework. *Frontiers in Psychology*, *10*, 2767. https://doi.org/10/ggfw2j

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs*. https://CRAN.R-project.org/package=BayesFactor

Murphy, J. H., & Banerjee, S. (2015). A caveat for the application of the critical cost efficiency index in induced budget experiments. *Experimental Economics*, *18*(3), 356–365. https://doi.org/10/f7nvgq

Nitsch, F. J., & Kalenscher, T. (2020). *Keeping a cool head at all times. What determines choice consistency?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/etyhx

Nitsch, F. J., Lüpken, L. M., Lüschow, N., & Kalenscher, T. (2021). *Inconsistently consistent: Rationality is not reliable* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/gd9zs

Nosofsky, R. M. (1986). Attention, Similarity, and the Identification-Categorization Relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57. https://doi.org/10/bjgj3w

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. In E. M. Pothos & A. J. Wills (Eds.), *Formal Approaches in Categorization* (pp. 18–39). Cambridge University Press. https://doi.org/10.1017/CBO9780511921322.002

Pedersen, T. L. (2019). *patchwork: The Composer of Plots* (R package version 1.0.0) [Computer software]. https://CRAN.R-project.org/package=patchwork

Plummer, M. (2003). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. *124*, 10.

Rieskamp, J., Busemeyer, J. R., & Mellers, B. A. (2006). Extending the Bounds of Rationality: Evidence and Theories of Preferential Choice. *Journal of Economic Literature*, *44*(3), 631–661. https://doi.org/10/dzqttm

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*(5), 356–374. https://doi.org/10/f4bx8f

RStudio Team. (2018). *RStudio: Integrated Development for R* (1.2.1335) [Computer software]. RStudio, Inc. ttp://www.rstudio.com/

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71. https://doi.org/10/cqz9wq

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703. https://doi.org/10/fpshqw

Steffener, J., Brickman, A. M., Rakitin, B. C., Gazes, Y., & Stern, Y. (2009). The Impact of Age-Related Changes on Working Memory Functional Activity. *Brain Imaging and Behavior*, *3*(2), 142–153. https://doi.org/10/cvh72j

Steffener, J., Habeck, C. G., & Stern, Y. (2012). Age-Related Changes in Task Related Functional Network Connectivity. *PLoS ONE*, *7*(9), e44421. https://doi.org/10/f38kfh

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Bayesian Inference for Kendall's Rank Correlation Coefficient. *The American Statistician*, *72*(4), 303–308. https://doi.org/10/gg7g2c

Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, 945–973. https://doi.org/10/b6s3sx

Varian, H. R. (1991). *Goodness-of-fit for revealed preference tests*. Department of Economics, University of Michigan Ann Arbor.

Varian, H. R. (2014). *Intermediate Microeconomics: A Modern Approach: Ninth International Student Edition*. WW Norton & Company.

Varian, H. R. (2006). Revealed preference. *Samuelsonian Economics and the Twenty-First Century*, 99–115. https://doi.org/10/cbssvr

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. https://doi.org/10/gfgt6p

Wallenius, J., Dyer, J. S., Fishburn, P. C., Steuer, R. E., Zionts, S., & Deb, K. (2008). Multiple Criteria Decision Making, Multiattribute Utility Theory: Recent Accomplishments and What Lies Ahead. *Management Science*, *54*(7), 1336–1349. https://doi.org/10/c2gpfx

Weber, E. U., Johnson, E. J., Milch, K. F., Chang, H., Brodscholl, J. C., & Goldstein, D. G. (2007). Asymmetric discounting in intertemporal choice: A query-theory account. *Psychological Science*, *18*(6), 516–523. https://doi.org/10/as9

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10/ggddkj

Wimmer, G. E., & Shohamy, D. (2012). Preference by Association: How Memory Mechanisms in the Hippocampus Bias Decisions. *Science*, *338*(6104), 270–273. https://doi.org/10/f39gxj

Wong, B. (2011). Points of view: Color blindness. *Nature Methods*, *8*(6), 441–441. https://doi.org/10/dxzwn5

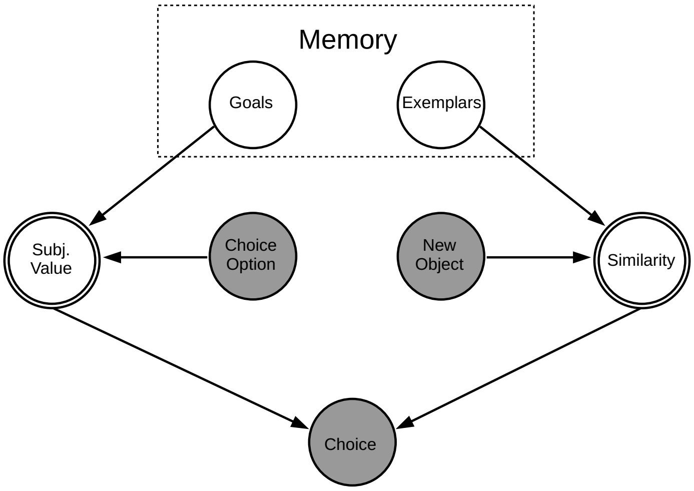
Zarahn, E. (2004). Positive Evidence against Human Hippocampal Involvement in Working Memory Maintenance of Familiar Stimuli. *Cerebral Cortex*, *15*(3), 303–316. https://doi.org/10/c2c23g

Zarahn, E., Rakitin, B. C., Flynn, J., & Stern, Y. (2006). Distinct spatial patterns of brain activity associated with memory storage and search. *NeuroImage*, *33*(2), 794–804. https://doi.org/10/b3w6zj

Zarahn, E., Rakitin, B. C., Flynn, J., & Stern, Y. (2007). Age-related changes in brain activation during a delayed item recognition task. *Neurobiology of Aging*, *28*(5), 784–798. https://doi.org/10/brq7gq

Figure 1

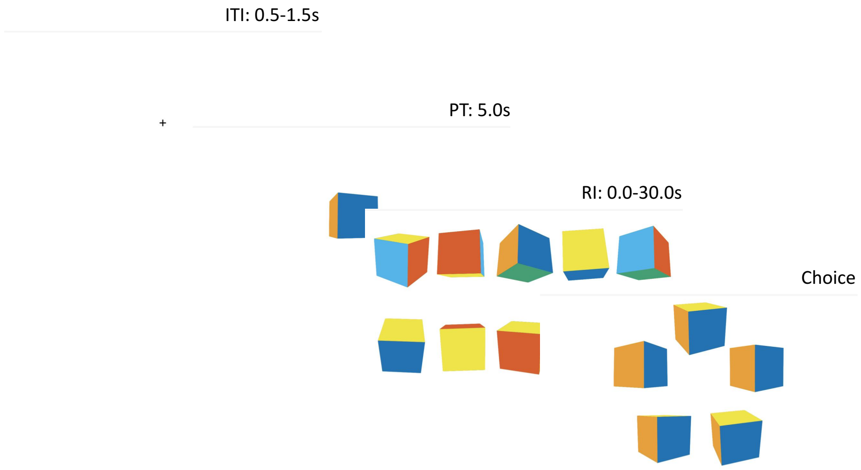
Mapping of concepts from GCM Model of Categorization to Query Theory



Note. Simplified graphical representation of the decision process in Query Theory (Weber et al., 2007) and the Generalized Context Model of categorization (Nosofsky, 2011). Nodes with dark background represent observed variables, nodes with white background represent latent variables. Nodes with single line borders represent stochastic variables, nodes with double line borders represent deterministic variables.

Figure 2

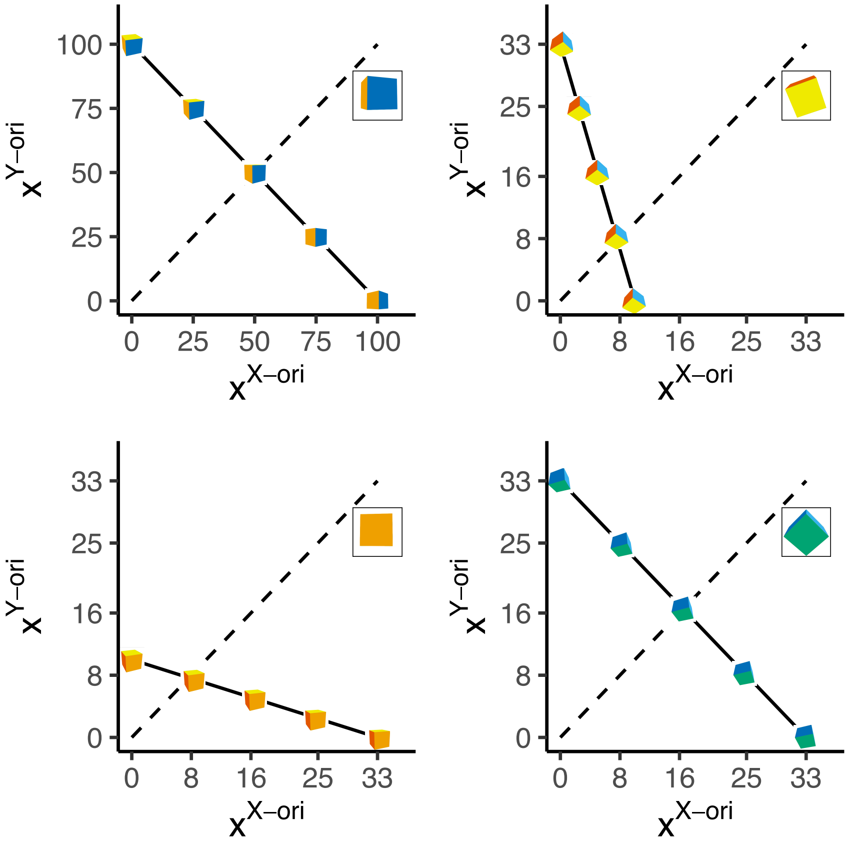
Timeline of a single choice trial



Note. From top left to bottom right: 1. The inter-trial interval (ITI) lasted 0.5 to 1.5 seconds. A fixation cross was presented in the middle of the screen. 2. During the presentation time (PT) an exemplar cube was presented for 5 seconds in the middle of the screen. 3. During the retention interval (RI) a mask of cubes in random orientations were presented. The retention interval was randomly selected from an interval 0 to 30 seconds and was fixed per participant per block. 4. After the retention interval, the choice set of 5 cubes with different orientations were presented. Each element of the choice set was presented equidistantly around the exemplar. The order of the choice set elements was randomized. Participants had to make a forced a choice on which among the choice set stimuli is most similar in its orientation to the exemplar.

Figure 3

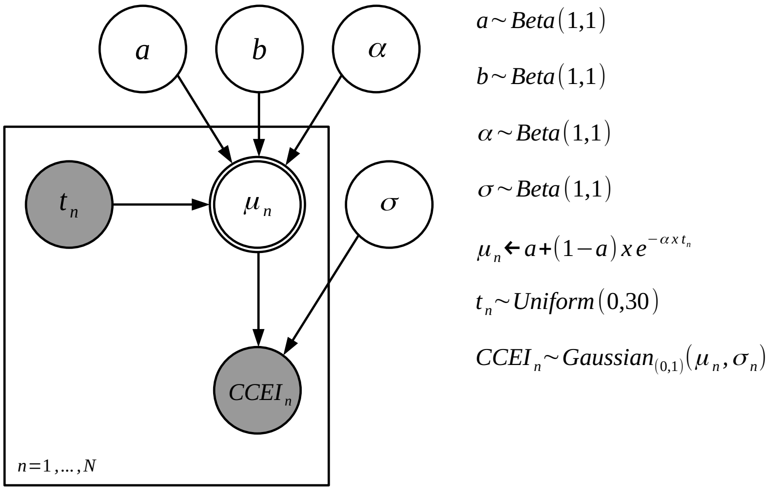
Construction of choice set of multi-attribute visual decision task from budget and prices



Note. Choice sets for four different exemplars and sets of prices. Exemplars are shown for each example in the square in the upper right corner of each panel. From top left to bottom right: $p_1 = (1,1), p_2 = (3,10), p_3 = (10,3), p_4 = (3,3)$. The size of the budget (set to $m = 100$) relative to the prices determines how similar the choice set stimuli are oriented to the exemplar overall. Hence, the choice set stimuli in the top left panel are overall more similarly oriented to their respective exemplar than the choice set stimuli in the bottom right panel. The price ratio of the attributes determines the trade-off ratio of the X- and Y-orientation. Hence, the choice set stimuli in the top right panel are generally more similarly oriented to their respective exemplar along the Y-axis and less similarly oriented along the X-axis compared to the bottom left panel and vice versa. Axiomatic choice theory proposes that subjective similarity increases as a function of how far a choice object is located to the top right (indicated by the dashed line).

Figure 4

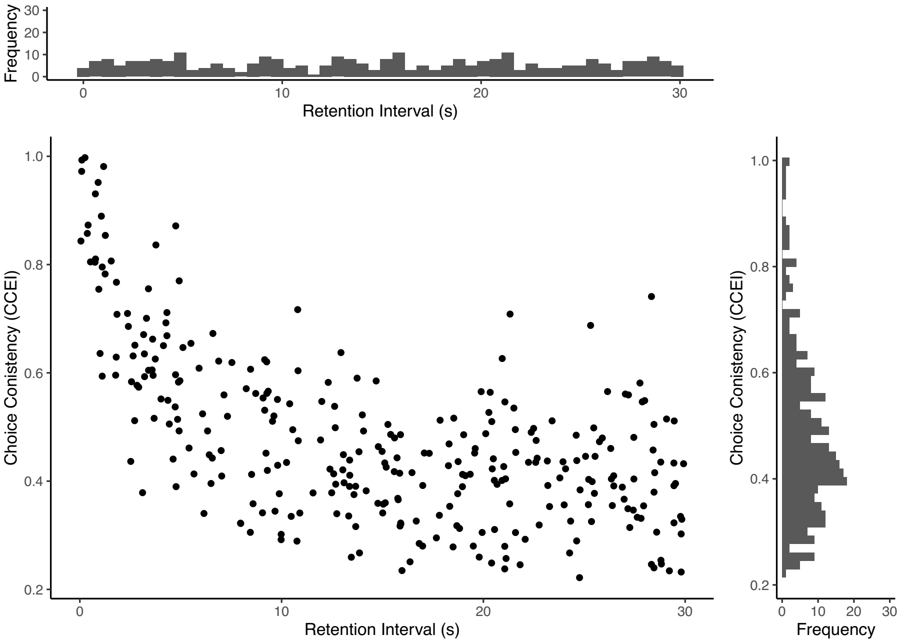
Graphical exponential model of the relationship of retention interval and choice-consistency.



Note. $n \in N$ denotes a single data point corresponding to a test block of a particular participant. t_n denotes the retention interval of a given observation. μ_n denotes the expected choice-consistency of a given observation. $CCEI_n$ denotes the observed choice-consistency of a given observation. The parameter a determines an asymptotical minimum level of choice-consistency after an infinite retention interval. The parameter b determines choice-consistency at $t = 0$, which allows for imperfect choice-consistency unconditional on time-dependent processes when $b < 1$. The parameter α determines the constant decreasing rate of consistency. We assumed that all parameter values are equally likely a priori. Nodes with dark background represent observed variables, nodes with white background represent latent variables. Nodes with single line borders represent stochastic variables, nodes with double line borders represent deterministic variables.

Figure 5

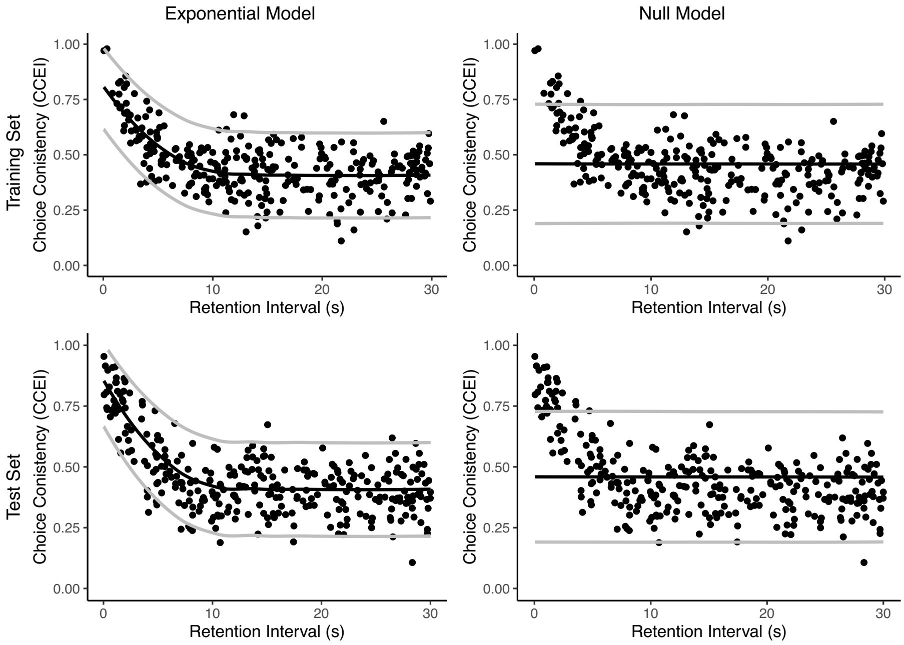
Scatterplot of retention interval and choice-consistency with histograms of marginal distributions.



Note. Data was simulated for 300 virtual participants using the exponential model with parameters $a = 0.4$, $b = 0.9$, $\alpha = 0.3$, $\sigma = 0.1$. The marginal distribution of the retention interval is, trivially, uniform. Importantly, the marginal distribution of choice-consistency is a right-tailed Gaussian, meaning a tail for large consistency values.

Figure 6

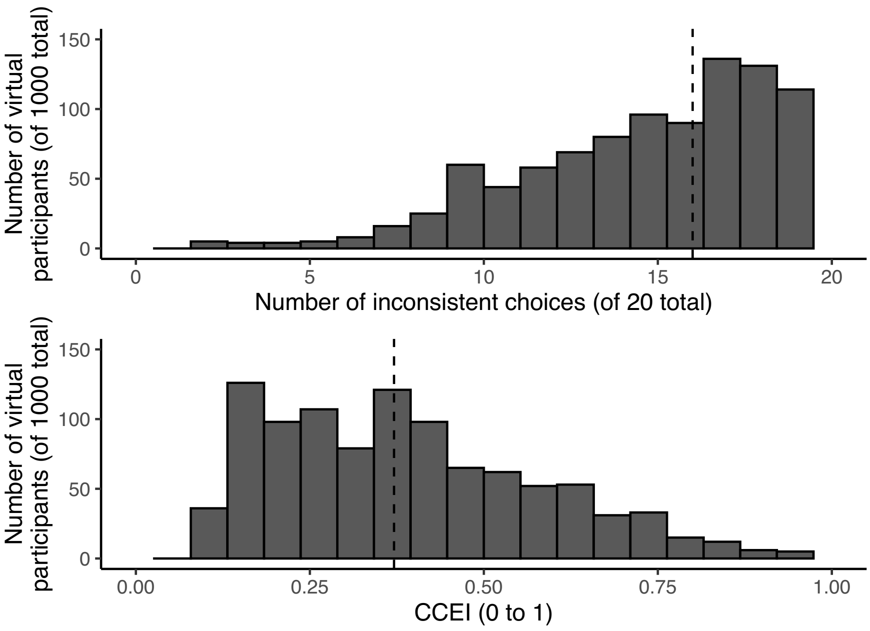
Scatterplots of retention interval and choice-consistency overlaid with posterior predictions



Note. The black line shows the median predictions, the grey lines show the 95% highest density intervals. Upper row shows the training set, bottom row shows the test set. Left column shows posterior predictions of the exponential model, right column shows posterior predictions of the null model. Training and test set were simulated for 300 virtual participants using the exponential model with parameters $a = 0.4$, $b = 0.9$, $\alpha = 0.3$, $\sigma = 0.1$. While the exponential model predicts the pattern of the data with relatively little uncertainty, the null model makes very vague predictions with possible values covering almost half of the variable space. Further, the null model does not predict the trend of the data for small retention intervals. Relatively to the training set performance of each model, the exponential model also generalizes slightly better to the test set.

Figure 7

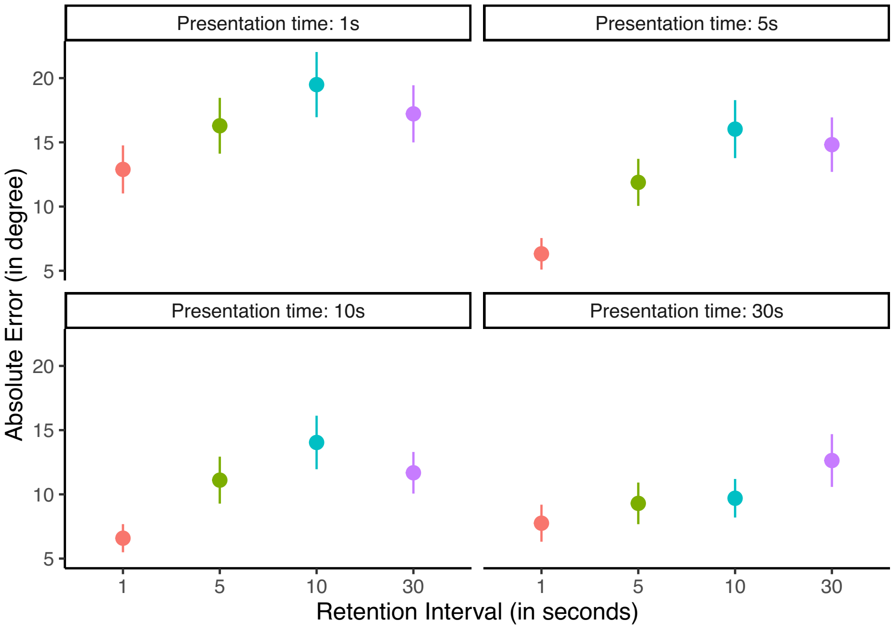
Histograms of choice-consistency of simulated random behavior



Note. We simulated a dataset of 1.000 virtual participants that made uniform random choices from 20 choice sets constructed as specified for our experiment (see Procedure). The upper panel shows the distribution of the number of inconsistent choices. 99% of the virtual participants committed at least one inconsistent choice. The median number if inconsistent choices (16) is indicated by the dashed vertical line. The lower panel shows the distribution of the CCEI. 99% of the virtual participants had a CCEI lower than 0.90. 97% of participants had a CCEI lower than 0.80. The median CCEI (0.37) is indicated by the dashed vertical line. Overall, our experimental task provides sufficient sensitivity to detect inconsistent choices. Note, that of all 1.000 virtual participants only a single one had a CCEI of 1. Importantly, this participant also did not violate the revealed preference axioms (0 inconsistent choices). We are, therefore, confident that our design also minimizes cost-efficient inconsistent choices which would undermine the sensitivity of the CCEI measure specifically (Murphy & Banerjee, 2015).

Figure 8

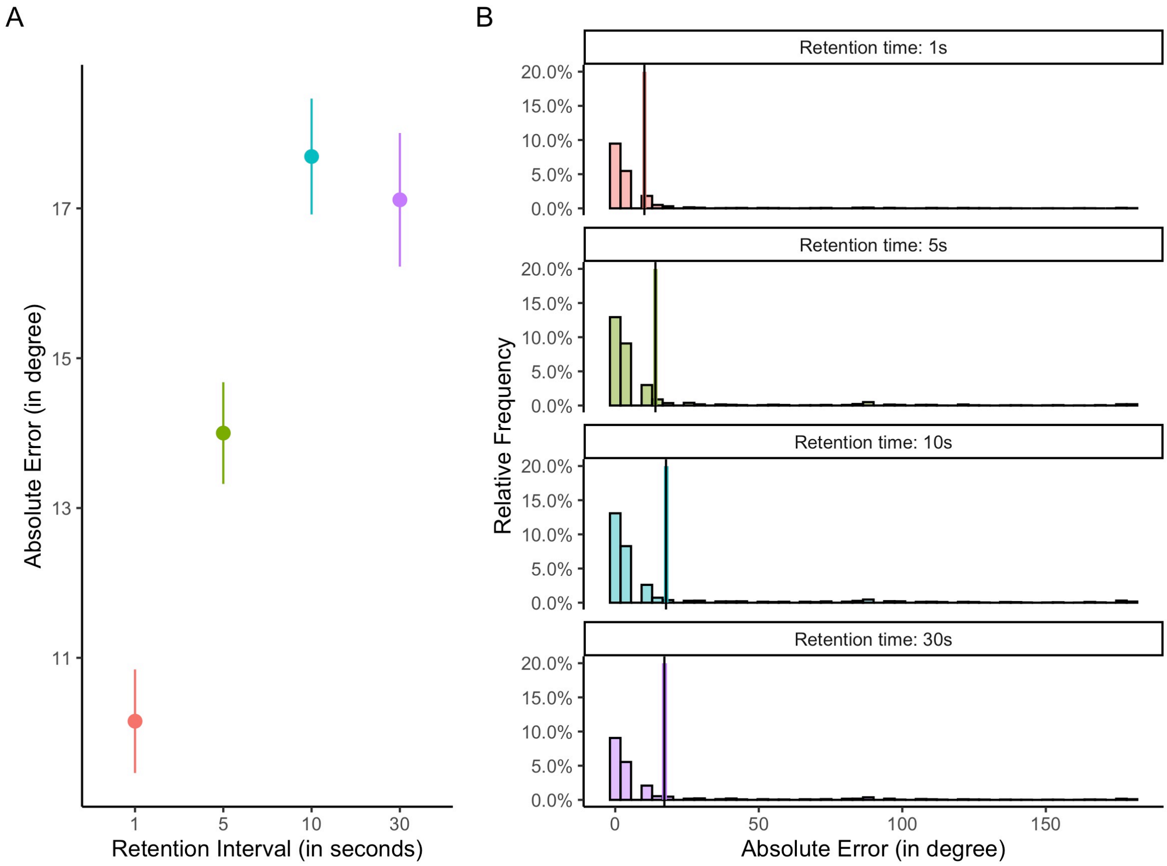
Results of pilot experiment



Note. The figure shows the point range (mean and standard error of the mean) for each cell of the two-factorial design. The absolute error in the reconstruction of the exemplar cube orientation is positively related to the retention interval and negatively related to the presentation time. There is definitive evidence against an interaction of both factors. Note, that the retention interval is not presented in scale.

Figure 9

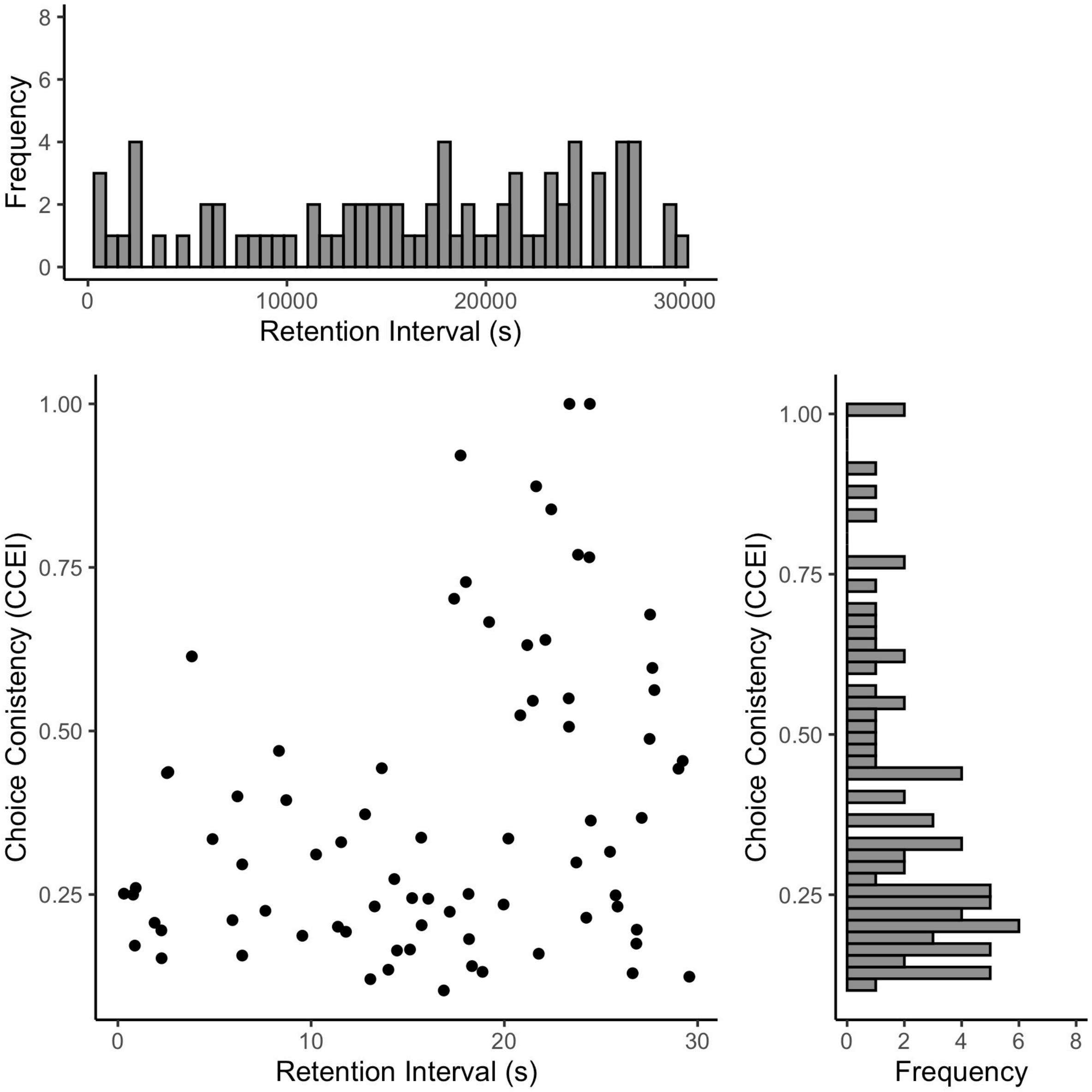
Manipulation check of the retention interval manipulation.



Note. Panel A shows the point range (mean and standard error of the mean) of the absolute reconstruction error for each retention interval level in the full data set (pilot and preregistered experiment, disregarding encoding time for the former). The absolute error in the reconstruction of the exemplar cube orientation is positively related to the retention interval (indicated by the evidence for the pilot data, the new data and the replication). Note, that the retention interval is not presented in scale. Panel B shows the histogram of the absolute reconstruction error for each retention interval level in the full data set (pilot and preregistered experiment, disregarding encoding time for the former). Vertical bars indicate the mean of the data, colored tiles indicate the standard error of the mean. It is eminent that the absolute error distribution has a strong positive skew.

Figure 10

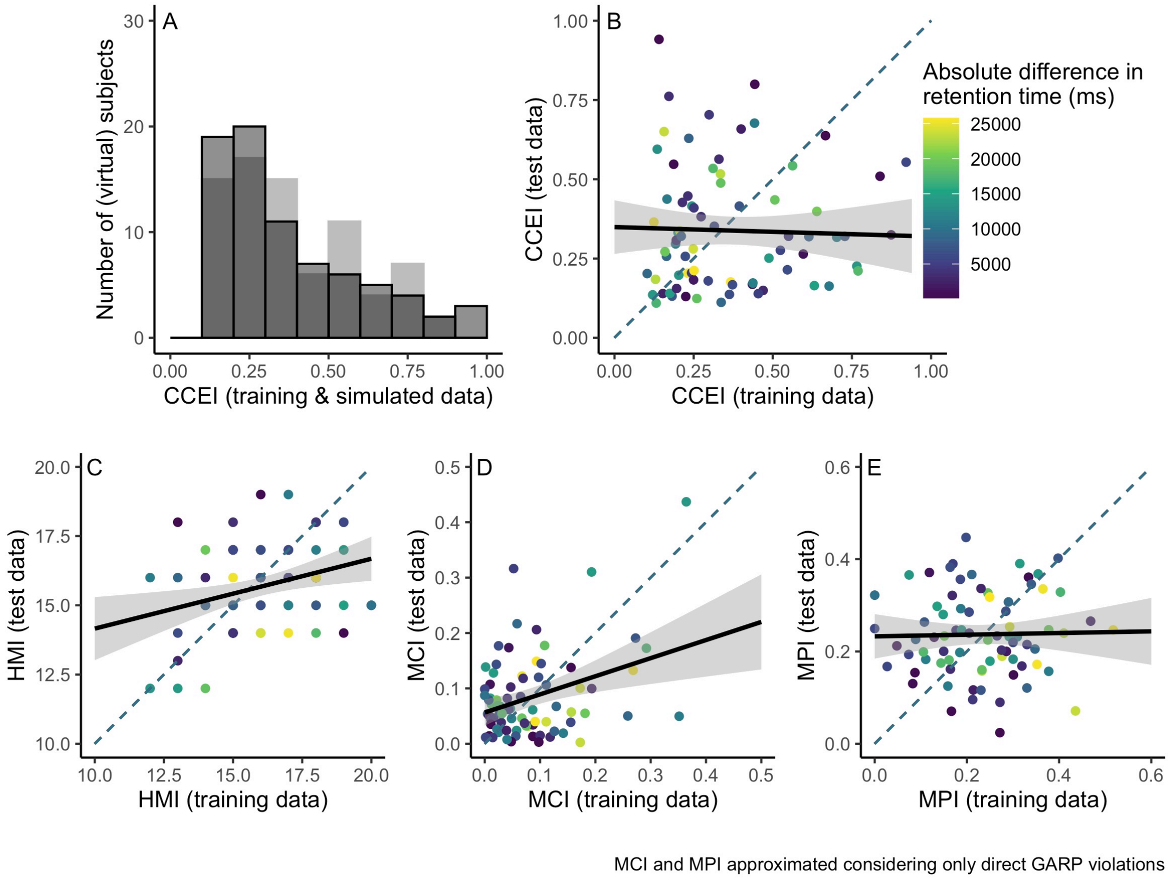
Scatterplot of the empirical retention interval and choice-consistency with histograms of marginal distributions.



Note. Trivially, the marginal distribution of the retention interval was uniform. Further, as expected, the marginal distribution of choice-consistency was a right-tailed Gaussian, meaning a tail for large consistency values. However, contrary to our hypothesis, the bivariate distribution plot revealed no negative relationship of CCEI and retention time.

Figure 11

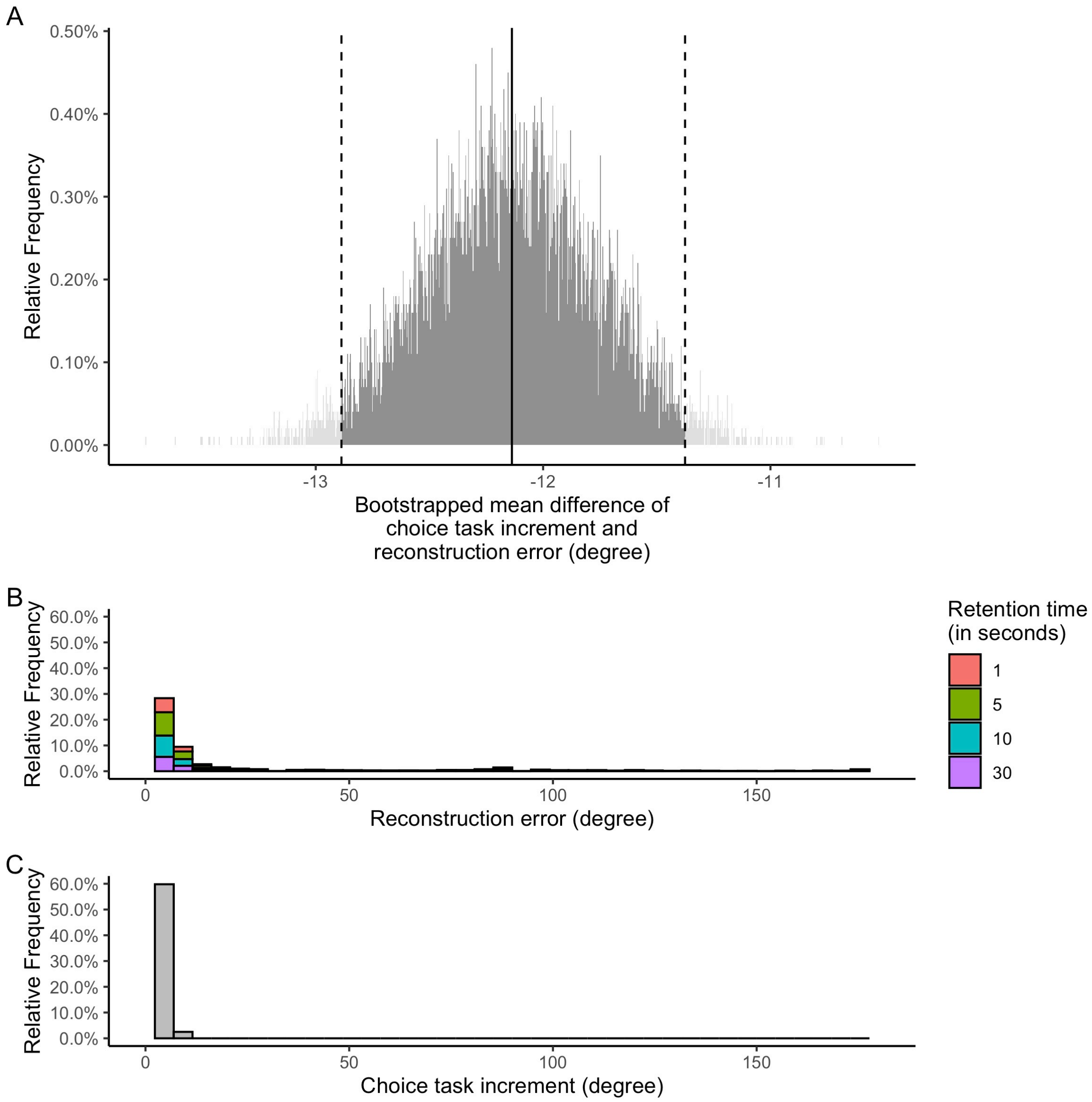
Empirical choice-consistency and reliablity



Note. Panel A shows the empirical distribution of the CCEI compared to an equally-sized subset of simulated random behavior. Choice-consistency, overall, was surprisingly low and not higher than for simulated random behavior. Panel B shows the test-retest reliability of the CCEI for training and test data, which was almost zero. Importantly, this was not driven by the absolute difference in retention time between both measurements. The lower panels (C, D, E) show similar patterns for three other consistency indices.

Figure 12

Exploration of task difficulty



Note. Panel A shows the bootstrapped distribution of the mean differences between the empirical absolute reconstruction error and choice task increment difference in the choice set (retrieved from task parameters). Depicted is the histogram of the bootstrap samples distribution (N = 10000). The continuous vertical line indicates the mean of the statistic of interest, the two dashed vertical lines indicate the 95% confidence interval. Results showed, that the mean increment difference was generally lower than mean reconstruction error, indicating a high difficulty of discriminating between the choice objects.

Table 1

Summary of statistical interpretation criteria for each hypothesis

Hypothesis	$BF \geq 10$	$BF \leq 0.1$	$0.1 < BF < 10$
H1: Inverse relationship of retention interval and choice-consistency	Strong support for inverse relationship	Strong support against inverse relationship	Inconclusive, larger N required
H2: Exponential model is supported more strongly by the data than null model	Strong support for exponential model	Strong support for null model	Inconclusive, larger N required
H3: The finding of H2 replicates to a new data set	Strong support for replication to a new dataset	Strong support against a replication to a new dataset	Inconclusive, larger N required

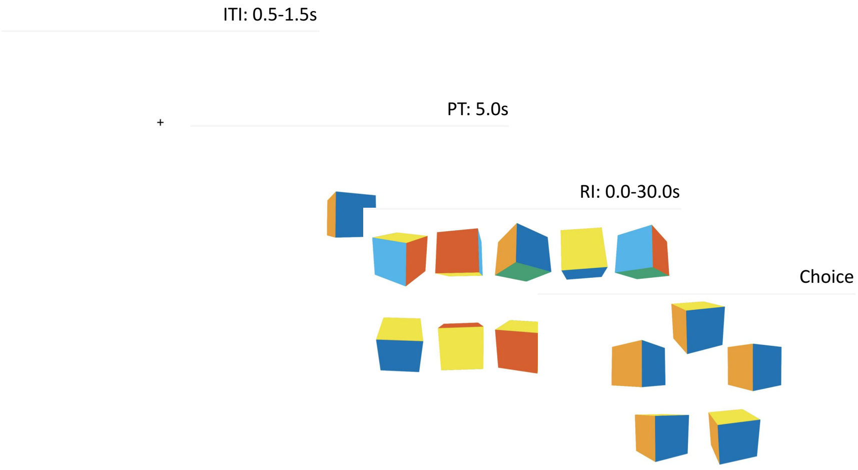
Appendix

English translation of instructions for MAVC Task

Dear participant,

In the following task you will be presented with a number of independent decision problems, that share a common format. Each decision problem starts with the presentation of colorful 3D cube. The cube will be presented for 5 seconds. Each side of the cube is identified by a unique color. Your task is to memorize the orientation of the cube as good as possible. After the presentation time of 5 seconds has passed, you will be presented with a visual mask of 10 similar cubes for up to 30 seconds. These cubes are irrelevant for the decision problem and you should not try to memorize their orientation. Finally, you will be present with 5 more cubes in different orientations which will be presented to you in circle. Your task is to select the one of those 5 cubes which has the most similar orientation to the cube which you have been presented with at the beginning of the decision problem. Before each decision problem you will be shortly presented with a fixation cross.

.



Carefully evaluate all 5 cubes and try to mentally rotate them until they match the memorized cube. Then select the cube which had to be mentally rotated the least. You have to decide for each decision problem. If you are unsure about your answer, follow your intuition. There are no wrong or correct answers. Before the task begins, please take a moment to familiarize with the colorful cube by inspecting the following animation. All cubes presented in the task will be exact copies of that cube but in different orientations.

[ANIMATION HERE]

Thank you for familiarizing with the colorful cube. You will now be presented with 10 practice decision problems. For these practice problems, the choice options will be presented 1 second after the cube that you have to memorize. Your answers for these practice decisions will not be recorded. Take your time to familiarize with the task.

You have successfully completed the practice decision problems. Do you have any questions or is there anything unclear about the task at hand? Then please raise your hand and consult with the experimenter.

If you have no further questions, then you can proceed now with the first test block. The test block consists of 20 decision problems. For all 20 decisions you will be assigned a retention interval of up to 30 seconds after the presentation of the initial cube during which you will see the irrelevant visual mask.

You have successfully completed the first test block. Take a moment to stretch your legs before continuing.

The next test block again consists of 20 decision problems. For all 20 decisions you will be assigned a different retention interval of up to 30 seconds after the presentation of the initial cube during which you will see the irrelevant visual mask.

Study 3 – Keeping a cool head at all times. What determines choice consistency?

Nitsch, F. J.*, & Kalenscher, T. (2020). *Keeping a cool head at all times. What determines choice consistency?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/etyhx

*Corresponding Author

CRediT Author Statement:

Felix Jan Nitsch: Conceptualization, Formal Analysis, Investigation, Writing - Original draft preparation, Visualization, Project Administration

Tobias Kalenscher: Conceptualization, Resources, Writing - Original draft preparation, Supervision, Funding Acquisition

Keeping a cool head at all times. What determines choice consistency?

Felix J. Nitsch and Tobias Kalenscher

Comparative Psychology, Heinrich-Heine-University Düsseldorf, Germany

Author Note

Felix J. Nitsch https://orcid.org/0000-0002-7832-7498

Tobias Kalenscher https://orcid.org/0000-0002-0358-9020

Correspondence concerning this article should be addressed to Felix J. Nitsch, Comparative Psychology, Heinrich-Heine-University Düsseldorf, 40225 Düsseldorf, Germany. Email: felix.nitsch@hhu.de

The research was supported by a grant from the Deutsche Forschungsgemeinschaft (Bonn, Germany) to Tobias Kalenscher (grant-ID: DFG-KA: DFG-KA 2675/4-3).

Abstract

Many rational choice theories posit that rational decision makers assign subjective values to all available choice options and choose the option with highest subjective value. Choice options are usually composed of multiple attributes, e.g. healthiness and taste in dietary choice or risk and expected returns in financial choice. These attributes have to be integrated into a single subjective value. Subjective value maximizing choice requires choice consistency, i.e. consistent weighing of the choice attributes across choices. However, empirical work suggests that perfect choice consistency is often violated, for example when decision makers weigh choice attributes differently across multiple decisions. Some researchers propose to extend certain bounds of rationality or to abandon the concept of rationality as adherence to consistency principles altogether. A more conservative stance assumes that perfect consistency can be violated by decision makers in practice, but that consistency principles still can explain large parts of behavior. In a review of the recent literature, we identify factors for compromised consistency relative to baseline conditions. Broadly, we distinguish between undynamic trait factors and fluid state factors. We find evidence for an influence of age, education, intelligence, and neurological status. In contrast, choice consistency appears to be relatively robust to the influence of sex, personality traits, cognitive load, sleepiness and blood alcohol levels. We conclude, that, according to the current state of the literature, only fundamental differences in decision makers, that is, trait differences, have a significant impact on choice consistency.

Keywords: Choice consistency, rationality, trait factors, state factors

Keeping a cool head at all times. What determines choice consistency?

Should you order one glass of champagne, a large *Stein* of Bavarian beer, or rather a non-alcoholic carafe of water? Choice theory posits that decision makers assign subjective values to all choice alternatives, rank order them according to their value, and choose the alternative with the highest value. The above example illustrates that choice alternatives are often composed of multiple attributes, e.g. magnitude (one glass, one carafe, one *Stein*), healthiness (alcoholic vs. non-alcoholic, sugar content etc.), subjective taste (the sourness of champagne vs. the bitterness of beer), cultural value or other components. Decision theory assumes that all of these attributes are integrated into a single subjective value. Importantly, the importance of taste as well as all other attributes is subjective and differs between individuals. Hence, in many decisions, there is no single best alternative. Instead, we have to deal with trade-offs of different choice attributes: In our diet we have to weigh taste, healthiness, magnitude or other factors. In financial decisions we have to weigh risk and expected returns, as well as the time until we can realize those returns. In such situations, our decisions depend on our individual preferences – so how can we measure decision making quality or rationality?

Because preferences are subjective, they are undisputable, an insight already noted by Immanuel Kant (1790). Positivist economic theories of rationality are agnostic about the specific direction of preferences. Instead they define rationality by placing bounds on preference structure, often in the form of consistency principles (Sugden, 1991). One example for a consistency principle is transitivity: If a decision maker chooses a glass of champagne over a glass of beer, and a glass of beer over a glass of water, she should also choose a glass of champagne over a glass of water. Decision makers who adhere to those consistency principles are assumed to choose their most preferred choice alternative, and, hence, make optimal choices. Table 1 provides an overview of different formalizations of choice consistency and their interdependence.

Hence, choice consistency is considered a hallmark of rational choice. That is, in order for a decision maker to always choose the best option according to their subjective preferences, she needs to choose consistently. On the flipside, an inconsistent decision maker systematically foregoes better options in choice situations. For example, if a decision maker chose champagne over beer, and beer over water, but *not* champagne over water, she would make an inconsistent choice. Rational choice theory then implies, that at least one chosen option along that choice chain would not be in accordance with the decision maker's true preferences. What is possibly worse, in a market situation, such a decision maker would continuously pay to swap water with beer, beer with champagne, and champagne with water again, thus, she would inevitably lose wealth and would, ultimately, be driven out of the market (this is the so-called *money pump* phenomenon).

Inferring the underlying preference structure of decision makers is non-trivial: Simply asking people about their preferences often gives conflicting results to actual choice behavior (MacDonald et al., 2009): People do not do what they say. Revealed preference theory (Houthakker, 1950; Samuelson, 1938; Varian, 1982) infers the underlying preference structure from behavior and allows for a theoretically sound test of choice consistency: that is, whether a decision maker has a definite structure of wants, acts cost efficient and acyclic (Afriat, 1973). Figure 1 provides schematic representation of a choice attribute weighing process compliant with revealed preference theory.

However, empirical work suggests that perfect consistency is often violated. Violations can occur when decision makers weigh choice attributes differently across multiple decisions. An inconsistent decision maker would assign variable weights to healthiness and taste in dietary choices, or to risk, expected returns and time of returns in financial choices, respectively: Choice inconsistency across multiple choices would imply that the different choice attributes (healthiness and taste; risk, expected returns, and time of returns) are differentially important to the decision maker across choices and may be considered to a further or lesser extent. For example, an inconsistent decision maker may mostly consider healthiness in one choice and mostly taste in another, leading her to choose the sugary option at one time, and the healthy option at another time. Under certain conditions, such inconsequential weighting of choice attributes can lead to violations of consistency principles, such as transitivity, or others (cf. table 1).

Some researchers propose to extend certain bounds of choice consistency (Rieskamp et al., 2006) or to abandon the concept of rationality as adherence to consistency principles altogether (Gigerenzer & Selten, 2002). This would have important theoretical implications — the rationality assumption is at the core of most economics, game theory and decision theory — and practical implications for the justification of policy decisions.

We are going to adopt a conservative stance of assuming that perfect consistency can be violated by decision makers in practice, but that consistency principles still can explain large parts of behavior. There are several sophisticated goodness of fit measures that quantify the degree to which choice behavior is consistent (Dean & Martin, 2016; Echenique et al., 2011; Heufer & Hjertstrand, 2015; Varian, 1993). The most prominent of these measures is Afriat's criticial cost efficiency index (Afriat, 1972, 1973; Varian, 1993), which has been used in many studies on choice consistency as defined by revealed preference theory (Andreoni & Miller, 2002; Banks et al., 2018; Bruyneel et al., 2012; Burghart et al., 2013; Cappelen et al., 2014; Castillo et al., 2017; Choi et al., 2007, 2014; Drichoutis & Nayga, 2017; Harbaugh et al., 2001; Kim et al., 2018; Lazzaro et al., 2016). The critical cost efficiency index is inspired by the fact that inconsistent choice behavior is not cost efficient. With given prices of choice options, the budget determines all affordable choice options. A revealed preference violation occurs, when the decision maker does not choose the most preferred alternative that was affordable given the budget, but selects another less preferred option. This can be interpreted as a waste of money, as the decision maker did not obtain the maximum subjective value for her money. The critical cost efficiency index denotes the minimal hypothetical reduction of the budget of the decision maker necessary, so that all more preferred but not chosen options become unaffordable when the revealed preference violation occurred. A critical cost efficiency index of 1 denotes perfect consistency: The budget does not need to be reduced. The index approaches zero as the behavior becomes more inconsistent and the budget needs to be reduced starkly to eliminate inconsistency. It is possible to impose (arbitrary) consistency bounds on choice behavior (Varian, 1993) or to benchmark choice behavior against simulated random choices

(Bronars, 1987). Simulated choices can also be used to determine the statistical power of a revealed preference test. An in our eyes sensible approach is comparing different experimental conditions regarding choice consistency or relating choice consistency to other variables in a correlational data set. This allows for identifying determinants of choice consistency without requiring perfect consistency under neutral conditions.

Scope of this literature review

Empirical work on rationality has been reviewed extensively by Rieskamp (2006). Specifically, Rieskamp focuses on whether or to which degree choice behavior can be reunited with axiomatic choice theory. In contrast, we will identify factors for compromised consistency relative to baseline conditions from recent literature instead of searching for a definite test of axiomatic choice theory as a concept.

We believe, that the term *rationality* bears different meanings in different fields and, therefore, is an umbrella term for theoretically different concepts. In order to avoid normative debates on the meaning of rationality, in the remainder of this review article, we will use the term *choice consistency* instead. While, as mentioned, choice consistency is considered a hallmark of rationality in many axiomatic choice theories and, therefore, bears important theoretical relevance, choice consistency is no necessary condition for rationality as defined by all fields of research. For example, Pham (2007) provided a multi-disciplinary review on the influence of emotion on rationality, specifically. While the scope of this review appears to be related to ours, Pham (2007) defines rationality without reference to revealed preference theory and, therefore, ultimately addresses a different question. We will interpret studies in terms of choice consistency even when the original publication uses the term rationality, as long as rationality is operationalized through revealed preference theory. For the purpose of this review, we are only considering choice data that are interpreted within the revealed preference framework, which we consider one of the most rigorous approaches to conceptualize and quantify choice consistency. We are by no means implying that other frameworks, or model-free analyses, yield less meaningful conclusions, but they do differ in their conceptual underpinning, and are, hence, difficult to compare (see below for a brief discussion of other consistency indices). We are going to differentiate between trait and state factors, as we find this to be the most intuitive structure of the current research on influence factors of choice consistency. Table 2 provides a concise summary of our findings.

Trait factors of choice consistency

In the next section we are going to review studies that investigated the effect of trait factors on choice consistency. As trait factors we consider temporally undynamic factors which are often subject to interindividual but less to intraindividual variability.

Age

Harbaugh et al. (2001) investigated food choice consistency in 7- and 11-year-old children and undergraduates. In their experimental task, participants had to choose one out of set of snack bundles for a total of 11 trials. Each snack bundle consisted of specific amounts of bags of chips and boxes of juice. Participants were ensured that at the end of the study one randomly selected trial would be paid out to them. In total, 31 second graders (7 years old) and 42 sixth graders (11 years old) and 55 undergraduates (21 years old) were included for the study. The study found a significant decrease in the number of inconsistent choices from second to sixth grade. However, this result did not hold for the critical cost efficiency index.

A similar design was deployed by Bruyneel et al. (2012) who investigated food choice consistency in kindergarteners, third graders and sixth graders with ages ranging from 5 to 12 years (on average 8 years). Similar to Harbaugh et al. (2001), participants had to choose one out of a set of snack bundles for a total of 9 trials. Each snack bundle consisted of specific amounts of grapes, tangerines and letter biscuits. Participants were ensured that at the end of the study one randomly selected trial would be paid out to them. In total, 39 kindergarteners, 31 third graders and 30 sixth graders were included for the study. The study found that belonging to the kindergartener group was a significant predictor of committing inconsistent choices. However, again, this result did not hold for the critical cost efficiency index at a significance level of 5%.

Echenique et al. (2011) investigated choice consistency in household-level food grocery purchases in a panel study of 494 households in an urban area of large mid-western US city between 1991 and 1993. The study found older households, with the average age of the spouses exceeding 65 years, were less consistent in their grocery purchases as measured by the money pump index, which is conceptually similar to critical cost efficiency index.

Choi et al. (2014) conducted an online study with 1,182 participants randomly selected from a survey designed to be representative of the Dutch population. In their experimental task, participants had to allocate a monetary endowment between two accounts. Both accounts were equally likely to be selected for payout: in half of the cases, the money in the first account would be selected for payout and in the other half, the second account would be selected. Importantly, the two accounts offered different relative returns. For example, the first account might offer double the returns of the second account or vice versa. The task challenged participants to balance expected value and variance of payout (risk). To maximize expected value, participants would need to allocate the complete monetary endowment to the account with higher relative returns. To minimize variance, participants would have to perfectly balance returns of the two accounts, so that both accounts would offer the same payout if they were selected. Except for rare cases, when both accounts offered equal relative returns, maximizing expected value and minimizing variance were mutually exclusive strategies. Participants solved a total of 25 trials. At the end of the study one trial was randomly chosen and one of the two accounts corresponding to the trial was selected for payout. The study found that an age above 50 was a significant predictor of scoring a lower critical cost efficiency index.

Brocas et al. (2019) investigated choice consistency of younger (age 18 – 34) and older (age 59 – 89) adults in a simple and a complex choice task. The simple choice task was an adaptation of Harbaugh (2001), with choice bundles consisting of two different types of snacks. The complex choice task was an adaptation of Bruyneel et al. (2012), with choice bundles consisting of three different types of snacks. Each trial consisted of a choice between only two bundles. Note, that this was accounted for in the consistency analysis, effectively compromising the power of the consistency test. The study found that, in the complex choice task, the older adults violated choice consistency significantly more often and more severely than younger adults (using their own original measures of choice consistency). The study did not find a difference between younger and older adults for the simple choice task.

Dean & Martin (2016) investigated choice consistency in household-level food grocery purchases in a panel study of 977 representative households in the Denver metropolitan area from 1993 to 1995. The study did not find a significant influence of age on choice consistency at a 5% significance level.

Chung, Tymula & Glimcher (2017) included 39 healthy adults over the age of 65 with functionally normal Mini-Mental State Examination scores for a study using whole-brain voxel-based morphometry. They used an experimental task conceptually similar to Harbaugh et al. (2001), where participants had to choose one out of a set of bundles of small presents. Each bundle consisted of specific amounts of two types of presents (e.g. sudoku books, cross word puzzles), which were previously rated to be desired by the participants. In their age restricted sample, they did not find that age was correlated with choice inconsistency. However, the authors largely attribute this to a statistical power problem.

Overall, there is evidence that both young and old age compromise choice consistency to some extent. However, the reported effects of young age predominantly show for the number of inconsistent choices. This measure of inconsistency does not distinguish between minor and major violations of consistency, which could lead to an overestimation of the effect size. More research is necessary especially for the effect of young age, to provide more robust effect size estimations and to investigate the causal mechanisms for potential age effects on choice consistency. Moreover, in most samples, the age of children is confounded with the amount of formal education they received. We, therefore, evaluate the role of education in making consistent choices in the following.

Education

Echenique et al. (2011) found that less educated households were significantly less consistent in their grocery purchases as measured by the money pump index (a consistency measure conceptually similar to the critical cost efficiency index).

Cappelen et al. (2014) addressed the question whether there is a development gap in choice consistency between the United States and Tanzania. They included students from one of the most reputable universities in each country, namely UC Berkeley and the University of Dar es Salaam. Nevertheless, the Tanzanian and the US subjects differed substantially in sociodemographic and economic backgrounds. The study used the experimental task of Choi et al. (2014), in which participants had to allocate monetary endowments between two risk accounts, for a total of 50 trials. 126 students from the US and 216 students from Tanzania were included. They found that students from Tanzania were significantly less consistent than students from the US as measured by the critical cost efficiency index. However, both samples displayed a high degree of choice consistency and the authors consider the reported differences in choice consistency as economically irrelevant.

Choi et al. (2014) also assessed the education level. They found that a high level of education was a significant predictor for scoring a higher critical cost efficiency index.

Kim et al. (2018) investigated the effect of education in a randomized-controlled design, by randomly granting a 1-year financial support program for education among 2812 female Malawi students. The program reduced absence and drop-out rates and increased scores in a qualification exam. The study measured choice consistency by lab-in-the-field experiments. Their task consisted of 20 choice trials in the risk domain, and additional 30 trials in the time domain. The risk domain task was similar to the task of Choi et al. (2014). In the time domain task, participants had to allocate money between two accounts with different payment dates. The study found, that receiving the education intervention was a significant predictor for a higher critical cost efficiency index in ninth graders for both choice domains, but not for tenth graders.

Dean & Martin (2016) found no significant influence of education on choice consistency in their dataset.

Banks, Carvalho & Perez-Arce (2018) assessed whether an educational reform in England affected choice consistency. They used the task of Choi et al. (2014) for a total of 25 trials. The *1972 Raising of the School Leaving Age Order* increased minimum allowed age to leave school from 15 to 16. The authors carried out an online panel study with 2,700 participants born between September 1, 1954 and August 31, 1960 who left school at age 16 or younger. They did not find an effect of the educational reform on choice consistency.

Overall, there is evidence for a positive relationship of education on choice consistency. However, a causal effect of education seems to be especially pronounced in younger children, as the educational intervention in Kim et al. (2018) did affect choice consistency of ninth graders but not tenth graders and Banks, Carvalho & Perez-Arce (2018) did not find an effect of an educational reform increasing the minimum allowed age to leave school from 15 to 16. Since both studies used a large sample, their null (sub-)findings are unlikely due to a lack of statistical power.

Intelligence

Bruyneel et al. (2012) also assessed mathematical, language and creative abilities of children through teacher ratings. They found that lower mathematical abilities predicted inconsistent choices and a lower critical cost efficiency index. Interestingly, the opposite pattern was found for language abilities: Higher language abilities predicted inconsistent choices and a lower critical cost efficiency index.

Choi et al. (2014) also assessed intelligence using the Cognitive Reflection Test (Frederick, 2005), which consists of three arithmetic riddles. Importantly, every riddle has an intuitive, but false answer. The cognitive reflection test is strongly correlated with measures of intelligence, but is not an intelligence test per se (Frederick, 2005). It has been shown to be related to decision making biases. Choi et al. found that the score of the Cognitive Reflection Test was significantly correlated with the critical cost efficiency index.

Brocas et al. (2019) also assessed intelligence using Raven’s Matrices Test (Raven, 1983), a visual pattern-detection task. They found that lower intelligence was significantly correlated with number and severity of choice inconsistencies in the complex choice task. However, as the authors note themselves, a causal interpretation is not possible due to their sampling plan: They found that younger adults scored significantly higher in the Raven’s Matrices Test than older adults, which confounds the result on intelligence with an age effect.

Harbaugh et al. (2001) investigated, in a subsample of 37 sixth graders (11 years old), whether choice consistency was related to performance in the Oregon Mathematics Problem Solving Assessment, an hour-long test of mathematical achievement for students. They did not find a significant correlation between both measures. However, due to the small size of the subsample the absence of evidence should not be interpreted as evidence of absence.

Overall, there is preliminary evidence for the relationship of intelligence and choice consistency. However, as Bruyneel et al. (2012) note, there are meaningful differences between different measures of intelligence which can explain divergent results. Future studies on the relationship of intelligence and choice consistency should deploy standardized and validated measures of intelligence.

Neurological status

Camille, Griffiths, Vo, Fellows, & Kable (2011) investigated the impact of ventromedial frontal lobe damage on choice consistency. The ventromedial frontal cortex is a brain region relevant for the representation of choice value (Levy & Glimcher, 2012). Therefore, it is a natural region of interest for investigating questions of choice consistency in the sense of subjective value maximization. Camille et al. (2011) used the experimental task of Harbaugh et al. (2001), substituting bags of chips for chocolate bars as a second type of snacks besides boxes of juice. The study included 9 participants with ventromedial frontal lobe damage and 22 age-, and education-matched controls. The study found that patients violated choice consistency significantly more often and scored a lower critical cost efficiency index than healthy controls.

Chung, Tymula & Glimcher (2017) also investigated the relationship of age-related grey matter brain atrophy and choice consistency. They found that a reduction of grey matter density in the ventro-lateral prefrontal cortex correlated significantly with higher frequency of inconsistent choices and a lower critical cost efficiency index. Furthermore, they found, in a meta-analysis, that the ventrolateral prefrontal cortex is often co-activated with regions relevant for choice value such as the ventro-medial prefrontal cortex, which substantiates the results of Camille et al. (2011).

Overall, there is preliminary evidence for the impact of impairments of value-related regions on choice consistency. Note, that the reviewed studies used only a relatively small sample size, so that conclusions should be drawn cautiously.

However, the importance of value-related regions for choice consistency finds additional support in the function brain imaging study of Kurtz-David et al. (2019). Kurtz-David et al. (2019) used the experimental task of Choi et al. (2014) for a total of 108 trials. The study found that trial-specific choice inconsistency is correlated to functional activity in the ventro-medial prefrontal cortex, the anterior and the posterior cingulate cortex. Trial-specific inconsistency was measured by the money metric index (Halevy et al., 2018), which is conceptually similar to the critical cost efficiency index, applying a leave-one-out procedure.

Sex and menstrual cycle

Choi et al. (2014) found that female sex was a significant predictor for a lower critical cost efficiency index, but their correlational design does not allow for a causal interpretation.

Lazarro et al. (2016) compared choice consistency of females over all phases of the menstrual cycle to male controls. They used the experimental task of Harbaugh et al. (2001), substituting bags of chips for chocolate cookies and boxes of juice for ounces of milk. The study included 39 females tested during all 4 phases of the menstrual cycle and 36 male controls. They found high levels of consistency (mean criticical cost efficiency index larger than 0.95) across all menstrual cycle phase and no difference to their male counterparts.

Overall, based on the reviewed studies there is no sufficient evidence for sex differences in choice consistency.

Personality traits

Cappelen et al. (2014) also assessed personality traits via the Big Five Inventory (John et al., 1991). None of the Big Five factors, namely Openness, Conscientiousness, Extraversion, Agreeableness and Neuroticism, significantly predicted choice consistency as measured by the critical cost efficiency index. Concludingly, there is no evidence that personality traits affect choice consistency. However, this conclusion is only based on a single study.

State factors of choice consistency

In the next sections we are going to review studies that investigated the effect of state factors on choice consistency. As state factors we consider temporally dynamic factors which are subject to inter- and intraindividual variability

Choice domain

Choice consistency has been tested in various domains, such as food choice (Bruyneel et al., 2012; Burghart et al., 2013; Camille et al., 2011; Harbaugh et al., 2001; Lazzaro et al., 2016), decisions under risk (Castillo et al., 2017; Choi et al., 2007, 2014; Drichoutis & Nayga, 2017; Kim et al., 2018), intertemporal choice (Andreoni & Sprenger, 2012; Chakraborty et al., 2017; Kim et al., 2018), altruistic decisions (Andreoni & Miller, 2002) and moral intuitions (Barbato et al., 2017). However, direct cross-domain comparisons have not been attempted to date. So far, Kim et al. (2018) is the only study that assessed choice consistency in two domains within the same sample. They do not report stark qualitative or quantitative differences in choice consistency with regard to the two different domains. Harbaugh et al. (2001) note, that they find more violations of consistent choice in food choice than Andreoni & Miller (2002) in social choice. However, the studies vary in their statistical power to detect inconsistent choices and sample composition, so that any differences cannot be clearly attributed to an effect of the choice domain.

Overall, it is unclear whether there are choice domain effects on choice consistency. Future studies should deploy choice tasks in multiple domains to ensure generalizability of results. Furthermore, a direct test of choice domain effects is required.

Cognitive Load

Drichoutis & Nayga (2017) investigated the effect of cognitive load on choice consistency. They used the experimental task of Choi et al. (2014) for a total of 60 trials. Furthermore, participants underwent 5 trials of a mental addition task, a mental multiplication task and a click-a-button task each. To manipulate cognitive load, participants had to solve an incentivized number-memorization task in parallel to the main task of each trial. Subjects in the high cognitive load treatment had to memorize 8-digit numbers, participants in the low cognitive load treatment had to memorize 1-digit numbers. The experiment included a total of 178 undergraduate participants from the Agricultural University of Athens, Greece. While the cognitive manipulation affected performance in the two mental arithmetic tasks it did not affect choice consistency in any considered measure, including the critical cost efficiency index. A post hoc power analysis showed that the study design was able to detect even small differences in choice consistency. Concludingly, there is no evidence that cognitive load affects choice consistency. However, this conclusion is only based on a single study.

Sleepiness

Castillo et al. (2017) investigated the effect of sleepiness on choice consistency. They used the experimental task of Choi et al. (2014) for a total of 50 trials. To manipulate sleepiness, participants were randomly assigned to an experimental session at a preferred time of the day relative to their diurnal preference or at a non-preferred time. The experiment included a total of 202 participants, with 115 participants doing the experiment at a preferred time. While the manipulation successfully affected sleepiness, there was no significant difference in choice consistency between both groups as measured by the critical cost efficiency index. Concludingly, there is no evidence that sleepiness affects choice consistency. However, this conclusion is only based on a single study.

Alcohol

Burghart et al. (2013) investigated the effect of blood alcohol concentration on choice consistency. They recruited participants from a bar in Manhattan, New York to conduct a field choice study. Blood alcohol concentration was measured by breath alcohol concentration. They used the experimental task of Harbaugh et al. (2001), substituting bags of chips for mini burgers and boxes of juice for dumplings. The study included a total of 101 participants with blood alcohol concentrations mostly uniformly distributed between 0.020% and 0.125%. Expected effects for these values range from minimal effects to major impairments of mental and physical control. The study found that blood alcohol concentration did not significantly predict choice consistency as measured by the critical cost efficiency index. Concludingly, there is no evidence that blood alcohol concentration affects choice consistency. However, this conclusion is only based on a single study.

Methodological caveats

Replications

A problem with drawing conclusions from the current literature on choice consistency is the lack of replications. While a few factors of choice consistency, such as age and education, have been the target of multiple studies, many of the reported effects have neither been replicated directly nor conceptually. Direct replications accumulate data to improve the precision of effect size estimates via meta-analysis. This can contribute to weeding out false-positive results (Nosek & Lakens, 2014). While conceptual replications are not best suited for validating a particular effect, they allow to abstract a phenomenon from its original operationalization (Nosek & Lakens, 2014). This is especially important for lab-based experimental studies, which often struggle to establish external validity.

Interpretation of effects and causality

Most studies reported here use frequentist statistics, that is p-value-based null hypothesis testing. However, p-values are inversely correlated with sample size (Lantz, 2013; Sullivan & Feinn, 2012). Since there are huge differences in the sample sizes of large-scale panel studies and lab-based experimental studies, comparative interpretation is difficult and not straight-forward. A stronger focus on effect sizes could, for example, contribute to understanding conflicting results on the effect of sex on choice consistency.

Another caveat lies in the interpretation of heterogeneous research designs. Some studies use a correlational approach for large representative datasets (Choi et al., 2014), sometimes using natural experiments to establish causality (Banks et al., 2018). These studies typically have a greater external validity, but do not contribute to understanding mechanisms and processes behind observed relationships. Other studies use pseudo-experimental approaches, often when the variable of interest cannot be manipulated (Bruyneel et al., 2012; Burghart et al., 2013; Cappelen et al., 2014; Chung et al., 2017; Harbaugh et al., 2001; Lazzaro et al., 2016). These studies essentially require the same caution in interpretation as correlational studies, as various confounds remain uncontrolled for. Furthermore, these studies often lack the large, representative samples of panel studies. Finally, there are studies deploying randomized controlled experimental designs (Castillo et al., 2017; Drichoutis & Nayga, 2017) which arguably are the gold-standard for establishing causality. However, these studies do not necessarily have strong generalizability beyond their original operationalization. Kim et al. (2018) deserve a special mention here, as their randomized controlled lab-in-the-field experiment promises both, high internal and external validity.

Process models

Few studies have tackled the question of how certain factors might influence choice consistency, which is especially relevant from a cognitive psychologist and neuroscientific point of view. A positive mention deserve the few studies investigating the neural underpinnings of choice inconsistency (Camille et al., 2011; Chung et al., 2017; Kalenscher et al., 2010; Kurtz-David et al., 2019), which suggest that value-encoding regions play a key role. Future studies should use a more theory-driven approach to identify moderators of choice consistency. Currently, it is unclear whether choice consistencies arise at the level of preference representations, choice value integration or behavior (trembling hand).

Choice of paradigm

The studies on choice consistency we reviewed in this articled are all heavily influenced in their paradigms by the two seminal papers of Harbaugh et al. (2001) and Choi et al. (2007), which are rooted in the framework of revealed preference theory. While these paradigms allow for a statistically powerful test of choice consistency, there might be specific effects on choice inherent to the construction of the paradigms. That is, precise critical cost efficiency estimation requires a specific way the decision problem is presented, e.g. in forms of graded choice bundles. It is possible that the very structure of the problem presentation is insensitive to relative changes in choice consistency, while the same decision problem framed differently might reveal stronger changes in relative inconsistency. For example, other paradigms similar to the one deployed by Tversky (1969) have borne different results. To enable meaningful interpretation, effects or their absence should, ideally, be robust across different paradigms.

Measure of consistency

Another concern is that critical cost efficiency, which is the basic measure of consistency in most of the findings reviewed here, is a compelling specification of choice consistency but not without alternatives. By design, its magnitude is solely determined by the strongest incidence of choice inconsistency in a given dataset. Therefore, it might not be sensitive to detect more subtle changes in choice consistency, which was for example the case in Harbaugh et al. (2001) and Bruyneel et al. (2012). Besides the other already mentioned goodness-of-fit measures for revealed preference theory, such as the money pump index (Echenique et al., 2011) or the money metric index (Halevy et al., 2018), there are several conceptually distinct operationalizations of choice consistency.

One approach is to parametrically estimate decision noise of a given choice model using a probabilistic choice rule (cf. Stott (2006) for an overview of probabilistic choice rules for binary choice). This parametric estimation of decision noise is favorable when there is a designated candidate model for the decision problem at hand and a general test of choice consistency is not necessary (Chumbley et al., 2014; Margittai et al., 2018; Sokol-Hessner et al., 2009).

Regenwetter et al. (2010) have proposed to conceptualize violations of choice consistency as the results of a mixture process of multiple for themselves consistent preference relations. Importantly, for each decision a consistent preference relation is probabilistically sampled from the collection of all consistent preference relations. Hence, their model treats choice inconsistency not as behavioral noise but as instable preferences.

Finally, the critical cost efficiency index is based on revealed preference theory and, thus, its validity is tied to the validity of revealed preference theory itself. Revealed preference theory has been conceptually and empirically criticized on several grounds (Arkes et al., 2016; Berg & Gigerenzer, 2010; Cason & Plott, 2014; Kőszegi & Rabin, 2007). Again, conceptual replications of the studies reviewed here using alternative measures of choice consistency are necessary to strengthen our confidence in the results.

Concluding Remarks

Here, we have reviewed recent literature to identify state and trait factors that determine the degree of choice consistency of decision makers as defined by revealed preference theory. While choice behavior of real-world decision makers systematically deviates from perfect consistency, it appears that the baseline degree of consistency seems to be relatively robust to trait and state factors. There is no or only limited evidence for an influence of sex and menstrual cycle (Lazzaro et al., 2016), personality traits (Cappelen et al., 2014), cognitive load (Drichoutis & Nayga, 2017), sleepiness (Castillo et al., 2017) or alcohol (Burghart et al., 2013) on choice consistency. There is evidence for an influence of age (Bruyneel et al., 2012; Choi et al., 2014; Harbaugh et al., 2001), education (Cappelen et al., 2014; Choi et al., 2014; Kim et al., 2018), intelligence (Bruyneel et al., 2012; Choi et al., 2014) and neurological status (Camille et al., 2011; Chung et al., 2017), but interpretation is often not possible without caution. Furthermore, there are also studies failing to find an effect of age, education or intelligence (Banks et al., 2018; Chung et al., 2017; Harbaugh et al., 2001). Overall, this suggests that choice consistency, apart from natural behavioral variability, seems to be a relatively robust trait of decision makers. Only fundamental differences in decision making ability, which might, for example, be age, education, and intelligence related, have an impact on choice consistency.

Future research on influence factors of choice consistency should replicate results directly and conceptually (in multiple paradigms), consistently report (standardized) effect sizes and develop a theoretical framework for the generative processes of choice inconsistency instead of endorsing a purely effect driven research agenda.

References

Afriat, S. N. (1972). Efficiency Estimation of Production Functions. *International Economic Review*, *13*(3), 568–598. JSTOR. https://doi.org/10.2307/2525845

Afriat, S. N. (1973). On a system of inequalities in demand analysis: An extension of the classical method. *International Economic Review*, 460–472. https://doi.org/10.2307/2525934

Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, *70*(2), 737–753. https://doi.org/10.1111/1468-0262.00302

Andreoni, J., & Sprenger, C. (2012). Estimating Time Preferences from Convex Budgets. *American Economic Review*, *102*(7), 3333–3356. https://doi.org/10.1257/aer.102.7.3333

Arkes, H. R., Gigerenzer, G., & Hertwig, R. (2016). How bad is incoherence? *Decision*, *3*(1), 20–39. https://doi.org/10.1037/dec0000043

Banks, J., Carvalho, L. S., & Perez-Arce, F. (2018). Education, Decision Making, and Economic Rationality. *The Review of Economics and Statistics*, *101*(3), 428–441. https://doi.org/10.1162/rest_a_00785

Barbato, M. T., Cosmides, L., Sznycer, D., & Guzmán, R. A. (2017). *Rational moral intuitions*. Human Behavior & Evolution Society.

Berg, N., & Gigerenzer, G. (2010). As-if behavioral economics: Neoclassical economics in disguise? *History of Economic Ideas*, 133–165. https://doi.org/10.2139/ssrn.1677168

Brocas, I., Carrillo, J. D., Combs, T. D., & Kodaverdian, N. (2019). Consistency in simple vs. Complex choices by younger and older adults. *Journal of Economic Behavior & Organization*, *157*, 580–601. https://doi.org/10.1016/j.jebo.2018.10.019

Bronars, S. G. (1987). The power of nonparametric tests of preference maximization. *Econometrica: Journal of the Econometric Society*, 693–698. https://doi.org/10.2307/1913608

Bruyneel, S., Cherchye, L., Cosaert, S., De Rock, B., & Dewitte, S. (2012). Are the Smart Kids More Rational? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2208412

Burghart, D. R., Glimcher, P. W., & Lazzaro, S. C. (2013). An expected utility maximizer walks into a bar... *Journal of Risk and Uncertainty*, *46*(3), 215–246. https://doi.org/10.1007/s11166-013-9167-7

Camille, N., Griffiths, C. A., Vo, K., Fellows, L. K., & Kable, J. W. (2011). Ventromedial Frontal Lobe Damage Disrupts Value Maximization in Humans. *Journal of Neuroscience*, *31*(20), 7527–7532. https://doi.org/10.1523/JNEUROSCI.6527-10.2011

Cappelen, A. W., Kariv, S., Sorensen, E., & Tungodden, B. (2014). Is There a Development Gap in Rationality? *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2432909

Cason, T. N., & Plott, C. R. (2014). Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing. *Journal of Political Economy*, *122*(6), 1235–1270. https://doi.org/10.1086/677254

Castillo, M., Dickinson, D. L., & Petrie, R. (2017). Sleepiness, choice consistency, and risk preferences. *Theory and Decision*, *82*(1), 41–73. https://doi.org/10.1007/s11238-016-9559-7

Chakraborty, A., Calford, E. M., Fenig, G., & Halevy, Y. (2017). External and internal consistency of choices made in convex time budgets. *Experimental Economics*, *20*(3), 687–706. https://doi.org/10.1007/s10683-016-9506-z

Choi, S., Fisman, R., Gale, D., & Kariv, S. (2007). Consistency and heterogeneity of individual behavior under uncertainty. *American Economic Review*, *97*(5), 1921–1938. https://doi.org/10.1257/aer.97.5.1921

Choi, S., Kariv, S., Müller, W., & Silverman, D. (2014). Who Is (More) Rational? *American Economic Review*, *104*(6), 1518–1550. https://doi.org/10.1257/aer.104.6.1518

Chumbley, J. R., Krajbich, I., Engelmann, J. B., Russell, E., Van Uum, S., Koren, G., & Fehr, E. (2014). Endogenous Cortisol Predicts Decreased Loss Aversion in Young Men. *Psychological Science*, *25*(11), 2102–2105. https://doi.org/10.1177/0956797614546555

Chung, H.-K., Tymula, A., & Glimcher, P. (2017). The Reduction of Ventrolateral Prefrontal Cortex Grey Matter Volume Correlates with Loss of Economic Rationality in Aging. *The Journal of Neuroscience*, *37*(49), 1171–17. https://doi.org/10.1523/JNEUROSCI.1171-17.2017

Dean, M., & Martin, D. (2016). Measuring rationality with the minimum cost of revealed preference violations. *Review of Economics and Statistics*, *98*(3), 524–534. https://doi.org/10.1162/REST_a_00542

Drichoutis, A. C., & Nayga, R. M. (2017). *Economic rationality under cognitive load*. Munich Personal RePEc Archive. https://mpra.ub.uni-muenchen.de/88192/

Echenique, F., Lee, S., & Shum, M. (2011). The money pump as a measure of revealed preference violations. *Journal of Political Economy*, *119*(6), 1201–1223. https://doi.org/10.1086/665011

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press. https://doi.org/10.7551/mitpress/1654.001.0001

Halevy, Y., Persitz, D., & Zrill, L. (2018). Parametric recoverability of preferences. *Journal of Political Economy*, *126*(4). https://doi.org/10.1086/697741

Harbaugh, W. T., Krause, K., & Berry, T. R. (2001). GARP for kids: On the development of rational choice behavior. *American Economic Review*, *91*(5), 1539–1545. https://doi.org/10.1257/aer.91.5.1539

Heufer, J., & Hjertstrand, P. (2015). Consistent subsets: Computationally feasible methods to compute the Houtman–Maks-index. *Economics Letters*, *128*, 87–89.

Houthakker, H. S. (1950). Revealed preference and the utility function. *Economica*, *17*(66), 159–174. https://doi.org/10.2307/2549382

John, O. P., Donahue, E. M., & Kentle, R. L. (1991). *The big five inventory—Versions 4a and 54*. https://doi.org/10.1037/t07550-000

Kalenscher, T., Tobler, P. N., Huijbers, W., Daselaar, S. M., & Pennartz, C. (2010). Neural signatures of intransitive preferences. *Frontiers in Human Neuroscience*, *4*, 49. https://doi.org/10.3389/fnhum.2010.00049

Kant, Immanuel. (1790). Analytik der ästhetischen Urteilskraft, §1 Das Geschmacksurteil ist ästhetisch. In Kant, Immanuel, *Kritik der Urteilskraft* (pp. 279–291).

Kim, H. B., Choi, S., Kim, B., & Pop-Eleches, C. (2018). The role of education interventions in improving economic rationality. *Science*, *362*(6410), 83–86. https://doi.org/10.1126/science.aar6987

Kőszegi, B., & Rabin, M. (2007). Mistakes in Choice-Based Welfare Analysis. *American Economic Review*, *97*(2), 477–481. https://doi.org/10.1257/aer.97.2.477

Kurtz-David, V., Persitz, D., Webb, R., & Levy, D. J. (2019). The neural computation of inconsistent choice behavior. *Nature Communications*, *10*(1), 1–14. https://doi.org/10.1038/s41467-019-09343-2

Lantz, B. (2013). The large sample size fallacy. *Scandinavian Journal of Caring Sciences*, *27*(2), 487–492. https://doi.org/10.1111/j.1471-6712.2012.01052.x

Lazzaro, S. C., Rutledge, R. B., Burghart, D. R., & Glimcher, P. W. (2016). The Impact of Menstrual Cycle Phase on Economic Choice and Rationality. *PLOS ONE*, *11*(1), e0144080. https://doi.org/10.1371/journal.pone.0144080

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: A neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038. https://doi.org/10.1016/j.conb.2012.06.001

MacDonald, E. F., Gonzalez, R., & Papalambros, P. Y. (2009). Preference inconsistency in multidisciplinary design decision making. *Journal of Mechanical Design*, *131*(3), 031009. https://doi.org/10.1115/1.3066526

Margittai, Z., Nave, G., Van Wingerden, M., Schnitzler, A., Schwabe, L., & Kalenscher, T. (2018). Combined effects of glucocorticoid and noradrenergic activity on loss aversion. *Neuropsychopharmacology*, *43*(2), 334. https://doi.org/10.1038/npp.2017.75

Nosek, B. A., & Lakens, D. (2014). Registered Reports: A Method to Increase the Credibility of Published Results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Pham, M. T. (2007). Emotion and Rationality: A Critical Review and Interpretation of Empirical Evidence. *Review of General Psychology*, *11*(2), 155–178. https://doi.org/10.1037/1089-2680.11.2.155

Raven, J. C. (1983). Manual for Raven's progressive matrices and vocabulary scales. *Standard Progressive Matrices*.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2010). Testing Transitivity of Preferences on Two-Alternative Forced Choice Data. *Frontiers in Psychology*, *1*. https://doi.org/10.3389/fpsyg.2010.00148

Rieskamp, J., Busemeyer, J. R., & Mellers, B. A. (2006). Extending the Bounds of Rationality: Evidence and Theories of Preferential Choice. *Journal of Economic Literature*, *44*(3), 631–661. https://doi.org/10.1257/jel.44.3.631

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71.

Sokol-Hessner, P., Hsu, M., Curley, N. G., Delgado, M. R., Camerer, C. F., & Phelps, E. A. (2009). Thinking like a trader selectively reduces individuals' loss aversion. *Proceedings of the National Academy of Sciences*, *106*(13), 5035–5040. https://doi.org/10.1073/pnas.0806761106

Stott, H. P. (2006). Cumulative prospect theory's functional menagerie. *Journal of Risk and Uncertainty*, *32*(2), 101–130. https://doi.org/10.1007/s11166-006-8289-6

Sugden, R. (1991). Rational Choice: A Survey of Contributions from Economics and Philosophy. *The Economic Journal*, *101*(407), 751–785. JSTOR. https://doi.org/10.2307/2233854

Sullivan, G. M., & Feinn, R. (2012). Using Effect Size—Or Why the P Value Is Not Enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Tversky, A. (1969). Intransitivity of preferences. *Psychological Review*, *76*(1), 31–48. https://doi.org/10.1037/h0026750

Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, 945–973. https://doi.org/10.2307/1912771

Varian, H. R. (1993). *Goodness of Fit for Revealed Preference Tests*.

Table 1

Different consistency principles and their implications

Consistency principle	Verbal description	Implies
First-Order Stochastic Dominance (FOSD)	If A is at least as good as B in all attributes and better in at least one attribute, then a decision maker should choose A over B.	GARP, IIA, SST, WST, Regularity
Generalized Axiom of Revealed Preference (GARP)	If a decision maker chooses A over B, and B over C then she should neither choose B or C over A, when A is strictly cheaper.	IIA, SST, WST, Regularity
Strong Stochastic Transitivity (SST)	If a decision maker chooses A over B with at least 50% probability and B over C with at least 50% probability, then she should choose A over C with a probability at least the larger of the probabilities of choosing A over B or C over B.	IIA, WST
Independence of Irrelevant Alternatives (IIA)	If a decision maker chooses A over C with a probability at least as great as that with which she chooses B over C, then she should choose A over D with a probability at least as great as that with which she chooses B over D.	SST, WST
Weak Stochastic Transitivity (WST)	If a decision maker chooses A over B with at least 50% probability and B over C with at least 50% probability, then she should choose A over C with at least 50% probability.	
Regularity	If a decision maker chooses options A and B each with a certain probability, then the addition of a third option C to the choice set may not increase these probabilities.	

Note. GARP: Generalized Axiom of Revealed Preferences, IIA: Independence of Irrelevant Alternatives, SST: Strong Stochastic Transitivity, WST: Weak Stochastic Transitivity.

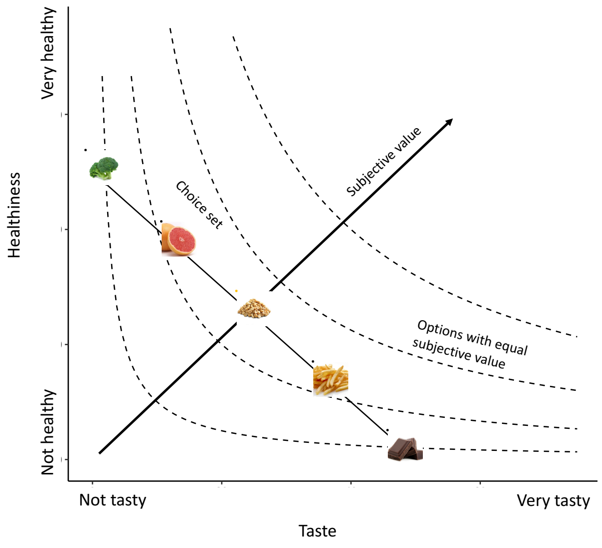
Table 2

Influence factors on choice consistency

Influence Factor	Trait or state?	Number of studies	Experimental designs	Studies reporting a significant effect
Age	Trait	6	Correlational, pseudo-experimental	4/6
Education	Trait	6	Correlational, pseudo-experimental, randomized controlled	4/6
Intelligence	Trait	4	Correlational, pseudo-experimental	3/4
Neurological status	Trait	2	Correlational, pseudo-experimental	2/2
Sex & menstrual cycle	Trait	2	Correlational, pseudo-experimental	1/2
Personality traits	Trait	1	Correlational	0/1
Cognitive load	State	1	Randomized controlled	0/1
Sleepiness	State	1	Randomized controlled	0/1
Alcohol	State	1	Pseudo-experimental	0/1

Figure 1

Schematic representation of consistent multi-attribute choice



Note. Imagine decision maker wants to decide on which snack to buy. She can choose between 5 different snacks (broccoli, grapefruit, peanuts, fries, chocolate), which all differ with regard to healthiness and taste. One option is very healthy but not tasty at all (broccoli). Another option is very tasty but not healthy at all (chocolate). Also, there are some compromise options (grapefruit, peanuts, fries) which are healthy and tasty to varying degrees. We call these 5 different snacks the choice set. Let us assume that the decision maker considers healthiness and taste equally important and prefers a compromise of health and taste over extreme options. When the preferences are well-defined like in our example, we can create collections of choice options which have the same overall subjective value to the decision maker (so-called *indifference curves*; dashed lines in the graph). Generally, collections of higher subjective value include options with higher healthiness and taste values. According to axiomatic choice theory, a decision maker should always choose the option from the choice set which is part of the collection of highest subjective value. In our example, the most preferred option is a perfect compromise of healthiness and taste (peanuts).

Study 4 – How robust is rational choice?

Nitsch, F. J.*, & Kalenscher, T. (2021). *How robust is rational choice?* [Preprint; accepted as poster at the annual meeting of the Society for Neuroeconomics]. PsyArXiv. https://doi.org/10.31234/osf.io/zv2m8

*Corresponding Author

CRediT Author Statement:

Felix Jan Nitsch: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization, Project administration

Tobias Kalenscher: Resources, Writing - Review & Editing, Supervision, Project administration, Funding acquisition

How robust is rational choice?

Felix J. Nitsch and Tobias Kalenscher

Comparative Psychology, Heinrich-Heine-University Düsseldorf, Germany

Author Note

*Corresponding author. Felix J. Nitsch, Comparative Psychology, Heinrich-Heine-University Düsseldorf, 40225 Düsseldorf, Germany. Email: felix.nitsch@hhu.de

The research was supported by a grant from the Deutsche Forschungsgemeinschaft (Bonn, Germany) to Tobias Kalenscher (grant-ID: DFG-KA 2675/7-1). The authors have no conflicts of interest with regard to the contents of this article. We want to thank Alina Keßler, Paul Kramer, and Marie Ludes for their assistance with the literature research and Paula Marie Klug for her assistance with polishing the p-curve disclosure table.

Abstract

Neoclassic economic choice theory assumes that decision-makers make choices as if they were rational agents. This assumption has been challenged over the last decades, yet systematic evidence aggregation beyond single experiments is still surprisingly sparse. Here, we asked how robust choice-consistency, as a proxy for rationality, is to endogenous and exogeneous factors. To this end, we conducted a systematic quantitative literature research, reviewing 5327 articles, identifying 44 as relevant that contained hypothesis tests on possible influence factors of choice-consistency. To assess the evidential value of any effect of such influence factors on choice-consistency, we conducted a p-curve analysis. Our results indicate that choice-consistency is affected by endogenous or exogeneous factors. This result holds for multiple testing procedures and a robustness check. However, due to the breadth of the contemporary research agenda, the lack of replications and the unavailability of original data in the field of choice-consistency, it is currently not possible to draw meaningful conclusions regarding specific influence factors. Despite this lack of specificity, our results implicate that people’s decisions might be a noisier and more biased indicator of their underlying preferences than previously thought. Hence, we provide systematic evidence for the wide-spread belief that rationality cannot be assumed unconditionally.

Introduction

Which career should I pursue? Which party should I vote for in the 2021 German federal elections? Decisions shape human lives and society, arguably, like no other psychological entity. The question of how to make good or *rational* decisions has puzzled philosophers, economists, and psychologists for centuries until today.

The predominant theory of rational choice is subjective utility maximization (SUM). In a nutshell, decision-makers are assumed to rank order all available choice options according to their subjective utility and select the one ranked highest. However, to this day it is not completely clear how, if at all, subjective utility values are represented (neuro-)psychologically (Hayden & Niv, 2020).

If treated as an non-psychological entity, subjective utility cannot be measured directly (Gul & Pesendorfer, 2008). Given this measurement problem, it is non-trivial to evaluate whether a specific decision was made for the option with maximum subjective utility and, therefore, was a good decision. Contemporary research generally uses tests on the choice structure of preferences. A milestone of 20th century economics lies in the proof that SUM requires consistency of the choice structure (Afriat, 1973; Houthakker, 1950; Samuelson, 1938; Varian, 1982): the choice of an option 1 over another, less expensive option 2 implies higher subjective utility of option 1.

Neoclassic rational choice theory has been criticized on several grounds (e.g. Arkes et al., 2016; Cason & Plott, 2014). However, a surge of recent publications pictures rationality as surprisingly robust (Nitsch & Kalenscher, 2020). A systematic and quantitative integration of these conflicting lines of research is currently missing, which we seek to provide here.

Specifically, we posed the research question of how robust choice-consistency, as a proxy for rationality, is to endogenous and exogeneous factors, such as age, drugs, education, emotions, financial status, intelligence, neurological status, personality, sex and gender, sleep deprivation, or stress. Importantly, we were interested in the influence of factors that may vary between participants, but are *assumed* to be constant within participants during the period of observation, so that covert responsive preference adaptation can be excluded as underlying source of overt inconsistency.

Results & Discussion

To answer how robust choice-consistency, as a proxy for rationality, is to endogenous and exogeneous factors, we conducted a systematic quantitative review of 5327 articles, identifying 44 research articles that contained hypothesis tests which addressed the influence of at least one of the predefined (see above), or related factors. To quantitatively aggregate the evidence in the literature, we conducted a P-curve analysis. P-curve analyses test how reliably a given effect is replicated in the literature by quantifying the evidential value and statistical power (Simonsohn et al., 2014, 2015).

The rationale behind this analysis is that, under the null hypothesis of no effect of endogenous or exogeneous factors on choice-consistency, the distribution of p-values of the effect of any such factor on choice-consistency in the published literature should follow a uniform distribution. On the other hand, in the presence of a true effect, the distribution of p-values should be positively and exponentially skewed. To account for the file drawer problem in scientific publications (Ioannidis et al., 2014; Nosek et al., 2013), the analysis only considers significant p-values ($p<.05$). Hence, to test our hypothesis, we considered independent significance tests for which the full test-statistic was reported or could be recalculated ($N=29$) and for which results indicated a significant effect ($p<.05$; $N=21$).

Descriptively, 17 of all 21 significant p-values (81%) fell in the lower half of the range of significant p-values ($p<0.025$; see figure 1, panel A). In line with this, our P-curve analysis indicated evidential value for that choice-consistency is affected by endogenous or exogeneous factors (see table 1; binomial test: $p=.0036$; Stouffer method: $Z=-10.97$, $p<.0001$ for full p-curve and $Z=-11.42$, $p<.0001$ for half p-curve). Further, we find no evidence that studies' evidential value is inadequate (see table 1; binomial test: $p=.9028$; Stouffer method: $Z=7.08$, $p>.9999$ for full p-curve and $Z=11.04$, $p>.9999$ for half p-curve). The statistical power estimate of the included studies amounted to 98% (90%-CI: 94%-99%).

Overall, it is important to point out that the breadth of the search for influence factors (roughly more than 20 different influence factors in 44 articles; see figure 2, panel B for a bibliographic analysis) stands in contrast to the severe lack of replications. Hence, it is currently not possible to draw meaningful conclusions regarding specific influence factors. Future research in the field should be careful to not only focus on finding novel influence factors but also validate the replicability of findings and find a common conceptual structure. In line with that, we should strive to rigorously make original data available to enable more efficient methods of data accumulation than the p-curve analysis utilized here.

Related to this, it is important to consider the heterogeneity of the investigated paradigms and influencing factors in the interpretation of the results: while we can infer that there are conditions under which rationality is compromised, this does not imply that rationality is compromised in the presence of every single influence factor investigated here.

The robustness of rationality has been a long-standing question in economics and psychology. Our results show that rationality cannot just be assumed for all decision-makers under all circumstances but instead endogenous and exogenous factors must be considered, even if these factors remain unchanged during the time period of observation. However, while our analysis unequivocally reveals that people's decisions might be a noisier and more biased indicator of their underlying preferences than previously thought, it remains unclear *what* it is exactly that makes them leave the path of rationality.

References

Afriat, S. N. (1973). On a system of inequalities in demand analysis: An extension of the classical method. *International Economic Review*, 460–472. https://doi.org/10.2307/2525934

Arkes, H. R., Gigerenzer, G., & Hertwig, R. (2016). How bad is incoherence? *Decision*, *3*(1), 20–39. https://doi.org/10.1037/dec0000043

Cason, T. N., & Plott, C. R. (2014). Misconceptions and Game Form Recognition: Challenges to Theories of Revealed Preference and Framing. *Journal of Political Economy*, *122*(6), 1235–1270. https://doi.org/10.1086/677254

Gul, F., & Pesendorfer, W. (2008). The case for mindless economics. *The Foundations of Positive and Normative Economics: A Handbook*, *1*, 3–42.

Hayden, B., & Niv, Y. (2020). *The case against economic values in the brain* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/7hgup

Houthakker, H. S. (1950). Revealed preference and the utility function. *Economica*, *17*(66), 159–174. https://doi.org/10.2307/2549382

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235–241. https://doi.org/10.1016/j.tics.2014.02.010

Nitsch, F. J., & Kalenscher, T. (2020). *Keeping a cool head at all times. What determines choice consistency?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/etyhx

Nosek, B., Spies, J. R., & Motyl, M. (2013). Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. https://doi.org/DOI: 10.1177/1745691612459058

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71.

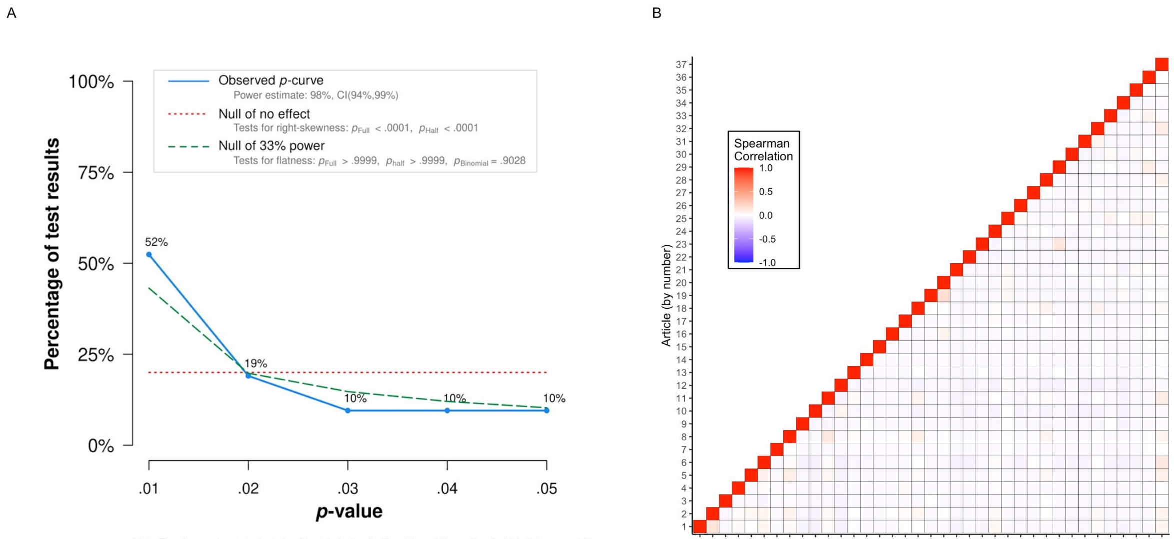
Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, 945–973. https://doi.org/10.2307/1912771

Figure 1

P-Curve and Co-Citation analysis



Note: The observed *p*-curve includes 21 statistically significant ($p < .05$) results, of which 17 are $p < .025$. There were 8 additional results entered but excluded from *p*-curve because they were $p > .05$.

Panel A created with P-Curve App 4.0

Panel A. P-curve analysis indicated evidential value for that choice-consistency is affected by endogenous or exogeneous factors. Further, we find no evidence that studies' evidential value is inadequate. The statistical power estimate of the included studies amounted to 98%. **Panel B.** The Article number refers to the bibliographic ID in the disclosure table (available online: https://doi.org/10.17605/OSF.IO/PAQ43). We were able to retrieve the full bibliographic record (including references) for 38 out of 44 articles from SCOPUS. The heatmap visualizes the overlap in referenced articles among the set of articles (Spearman rank correlation) as a rough quantitative measure of conceptual connectedness. In alignment with a qualitative coding of the investigated influence factors (see disclosure table), this quantitative analysis suggests little overlap in the conceptual structures of the included articles, undermining the breadth of the current research program on influence factors of choice consistency.

Table 1

	Binomial test (Share of results p<.025)	Continuous Test (Stouffer Method)	
1) Studies contain evidential value. (Right skew)	p=.0036	Full p-curve (p<.05)	Half p-curve (p<.025)
		Z=-10.97, p<.0001	Z=-11.42, p<.0001
2) Studies' evidential value, if any, is inadequate. (Flatter than 33% power)	p=.9028	Z=7.08, p>.9999	Z=11.04, p>.9999
Statistical Power of tests included in p-curve (correcting for selective reporting)	Estimate: 98% 90% Confidence interval: (94% , 99%)		

How robust is rational choice? – Supplemental Material

Felix J. Nitsch and Tobias Kalenscher

Comparative Psychology, Heinrich-Heine-University Düsseldorf, Germany

Methods

Models of choice consistency

Neoclassic economic choice theory assumes that decision-makers behave *as if* maximizing the subjective value or *utility* obtained with their choices. Subjective utility can be understood as the integrated hedonic value of a choice option and is by definition a latent variable. Hence, it is non-trivial to evaluate whether a specific decision was made for the option with maximum subjective utility and, therefore, was a good decision. If treated as an non-psychological entity, subjective utility cannot be measured directly (Gul & Pesendorfer, 2008). This measurement problem poses a significant challenge to empirical applications of utility theory. Contemporary research generally uses tests on the choice structure of preferences. A milestone of 20th century economics lies in the proof that subjective utility maximization (SUM) requires consistency of the choice structure (Afriat, 1973; Houthakker, 1950; Samuelson, 1938; Varian, 1982).

How can choice consistency be operationalized? For the scope of this article, we did not restrict our operationalization of choice consistency to a single model. This reflects the state of literature as well where a variety of choice consistency models is used. In the following we will give an overview of the most important model classes.

Perhaps the simplest model of choice consistency is choice variability. That is, we define consistency by the degree with which decision makers show identical choices for identical choice problems. According to this model, a decision maker is consistent if, and only if the choice of option A over option B, everything else equal, implies that option B is not chosen over option A. Choice consistency in this sense can be quantified by the relative choice frequencies in identical choice problems. A decision maker always choosing option A over option B is considered maximally consistent. A decision maker choosing options A and B with equal frequency is considered maximally inconsistent.

The advantage of this operationalization lies in its strict parsimony. It requires few theoretical assumptions and can be applied to virtually any domain and presentational format of choice, thus, being quite general. A disadvantage is, arguably, that only very little information about the choices and their context is used, as the model does not make any predictions on the relation of non-identical choice problems. Further, it is unclear what can be defined as an identical choice problem. It is virtually impossible to keep all variables constant across two choice problems, even in the context of a laboratory experiment and much less under realistic conditions. As the model does not make any assumptions about which variables are relevant, the analysis usually relies on *ad hoc* assumptions of the researchers, which might be difficult to make for complex choice problems. Most importantly, however, the model does not differentiate between inconsistency and indifference.

A generalization of the choice variability model of consistency is choice transitivity. According to this model, we define consistency by the degree to which decision makers' choices adhere to the mathematical property of transitivity: A decision maker is consistent if and only if the choice of option A over an intermediate option B and the choice of intermediate option B over option C, everything else equal, implies that option C is not chosen over option A. More generally, we can allow for multiple intermediate options B1, B2, … BN for which a fully connected chain of choices from options A to C can be identified. We can indicate the number of allowed intermediate options as the degree of transitivity, so that allowing for only one immediate option would be indicated as transitivity of the first degree. The choice variability model can then be reformulated as transitivity of degree zero.

An advantage of the transitivity over the variability model is that not only identical, but all choice problems are considered for the evaluation of choice consistency. Thus, it uses much more of the information in the data. Still, it can be applied to most domains and presentational formats of choice. Commonly, a threshold is imposed on the relative choice frequencies for identical choice problems, so that an option A is only considered not indifferent to option B for the decision maker, if it is chosen in a clear majority or minority of cases. However, the aggregation of multiple choices in combination with the problem of defining what are identical choice problems can lead to paradoxical conclusions. For example, aggregated choices might appear intransitive even when they consist of fully transitive subsets (i.e. preference states) and vice versa (also called Condorcet paradox; see Regenwetter & Davis-Stober, 2012). Again, as the model, too, does not make any assumptions about which variables are relevant for evaluating identity of choice problems, the analysis usually relies on *ad hoc* assumptions of the researchers, which might be difficult for complex choice problems.

The even more sophisticated revealed preference theory can be considered the dominant theory of consistent and, thus, rational choice in neoclassic economics. In a way, it can be seen as an extension of the choice transitivity model. In a nutshell, it maintains the requirement of transitivity, but takes a different approach to differentiating preference from indifference than only looking at relative choice frequencies. Instead, an additional assumption is made: The choice of an option A over another option B does not preclude indifference as long as option B is normatively better than option A in at least one relevant variable. Put in a more standard economic setting with positively priced goods, the choice of an option A over another option B does not preclude indifference as long as option B is at least as expensive as A (note, however, that revealed preference theory can be generalized to non-economic choice problems; see Nitsch & Kalenscher, 2020b). While revealed preference analysis also requires, but does not impose, a definition of relevant variables, it is much more explicit about this fact than the two previously discussed models. Extensions of revealed preference theory further allow for the quantification of inconsistency, for example via the threshold of just noticeable differences in the variables of interest (Afriat, 1972; Dziewulski, 2018).

A common characteristic of the three aforementioned models of choice consistency is that they impose restrictions on observed choice behavior, but make no or very little assumptions on the cognitive processes generating the behavior. Generally, this theoretical parsimony is bought dearly by requiring *ad hoc* assumptions of what are relevant variables by researchers in scientific practice.

Generative models take an orthogonal approach to this, by making specific assumptions about generative processes, from which variables of interest can be deducted. The advantage of these models is that they are well-specified and usually do not require auxiliary assumptions beyond their specification. Often, these models contain a nuisance parameter that captures how well actual behavior is aligned with the model predictions. This nuisance parameter can then be interpreted as a measure of choice consistency.

A problem with generative models is that the assumed generative process might be (severely) mis-specified. Therefore, the nuisance parameter does always include both, choice inconsistency and model specification error, and critically hinges on model validity. In conclusion, each discussed model is probably more desirable than the others in at least one relevant aspect.

Note, that there also models that do not require choice consistency in the sense of economic theory (and, thus, were not considered here) but which still imply goal-directed, non-random behavior (i.e. heuristics; Gigerenzer & Selten, 2002).

Search strategy and data extraction

To answer whether and which influence factors compromise choice consistency, we aimed to include literature on a wide range of candidate factors, which we identified based on a previous, non-systematic review of the literature (Nitsch & Kalenscher, 2020a). These candidate factors were age, drugs, education, emotions, financial status, intelligence, neurological status, personality, sex and gender, sleep deprivation, and stress. In May 2020, we queried four literature databases for the fields of psychology and economics (PubPsych, PsycInfo, EconBiz, Web of Science) resulting in 5327 articles found (3064 without duplicates). Articles retrieved from the databases were considered eligible if and only if they addressed the influence of at least one of the predefined factors (or related factors) on choice consistency. Eligibility was verified iteratively in two steps. For the first eligibility check we only considered the title and abstract of the articles, yielding 201 articles.

Unfortunately, for 37 of those articles we could not retrieve the fulltext. For the remainder of 164 articles we checked eligibility by reading the fulltext, yielding a final sample of 44 articles. The final amount of articles yielded by the systematic search is aligned with preliminary scoping of the literature. The low inclusion rate was driven by the precision of our research question. For these articles, we identified all investigated influence factors of choice consistency and extracted the p-value of the corresponding hypothesis test according to the P-Curve User-Guide (Simonsohn, Nelson, et al., 2015). If multiple tests were reported for a single influence factor, we selected the first significant test reported in manuscript for the main analysis and the second significant test reported for the robustness check. If no significant test was reported we included the first test reported overall. As the latter step was partly subjective, two people conducted this procedure independently, of whom one was blind to the research question of this article. Any discrepancies were resolved by discussion. Due to the nature of our analysis, we excluded articles that did not report p-values. Exact p-values were recalculated from the reported test statistics. If this was not possible, we excluded the article from the analysis. The full disclosure table is available online: https://doi.org/10.17605/OSF.IO/PAQ43If.

Data Analysis

P-Curve analysis was conducted using the P-Curve App (Simonsohn et al., 2014; Simonsohn, Simmons, et al., 2015). Articles and test-statistics were included following the P-Curve User Guide (Simonsohn, Nelson, et al., 2015; see disclosure table). The bibliometric analysis was conducted using Bibliometrix (Aria & Cuccurullo, 2017). The full output of the P-Curve App and the R-Code for the bibliometric analysis area available online: https://doi.org/10.17605/OSF.IO/PAQ43If.

Robustness Check

The results for the robustness check (see Search Strategy and Data Extraction) confirm our main results. Descriptively, 17 of all 21 significant p-values (81%) fell in the lower half of the range of significant p-values ($p<0.025$). In line with this, our P-curve analysis indicated evidential value for that choice-consistency is affected by endogenous or exogeneous factors (binomial test: p=.0036; Stouffer method: Z=-10.52, p<.0001 for full p-curve and Z=-10.93, p<.0001 for half p-curve). Further, we find no evidence, that studies' evidential value is inadequate (see table 1; binomial test: p=.9028; Stouffer method: Z=6.65, p>.9999 for full p-curve and Z=10.58, p>.9999 for half p-curve). The statistical power estimate of the included studies amounted to 97% (90%-CI: 92%-99%).

References

Afriat, S. N. (1972). Efficiency Estimation of Production Functions. *International Economic Review*, *13*(3), 568–598. JSTOR. https://doi.org/10.2307/2525845

Afriat, S. N. (1973). On a system of inequalities in demand analysis: An extension of the classical method. *International Economic Review*, 460–472. https://doi.org/10.2307/2525934

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Dziewulski, P. (2018). *Just-noticeable difference as a behavioural foundation of the critical cost-efficiency index*. https://www.economics.ox.ac.uk/materials/working_papers/4603/848-dziewulski.pdf

Gigerenzer, G., & Selten, R. (2002). *Bounded rationality: The adaptive toolbox*. MIT press. https://doi.org/10.7551/mitpress/1654.001.0001

Gul, F., & Pesendorfer, W. (2008). The case for mindless economics. *The Foundations of Positive and Normative Economics: A Handbook*, *1*, 3–42.

Houthakker, H. S. (1950). Revealed preference and the utility function. *Economica*, *17*(66), 159–174. https://doi.org/10.2307/2549382

Nitsch, F. J., & Kalenscher, T. (2020a). *Keeping a cool head at all times. What determines choice consistency?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/etyhx

Nitsch, F. J., & Kalenscher, T. (2020b). *Influence of memory processes on choice consistency.* https://doi.org/10.17605/OSF.IO/AKGT9

Regenwetter, M., & Davis-Stober, C. P. (2012). Behavioral Variability of Choices Versus Structural Inconsistency of Preferences. *Psychological Review*, *119*(2), 408–416. https://doi.org/10.1037/a0027372

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

Simonsohn, U., Nelson, L., & Simmons, J. (2015). *Official User-Guide to the P-curve*. 6.

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, 945–973. https://doi.org/10.2307/1912771

Study 5 – Inconsistently consistent: Rationality is not reliable

Nitsch, F. J.*, Lüpken, L. M., Lüschow, N., & Kalenscher, T. (2021). *Inconsistently consistent: Rationality is not reliable* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/gd9zs

*Corresponding Author

CRediT Author Statement:

Felix Jan Nitsch: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization, Supervision, Project Administration

Luca Marie Lüpken: Conceptualization, Software, Investigation, Writing - Review & Editing

Nils Lüschow: Software, Writing - Review & Editing

Tobias Kalenscher: Conceptualization, Writing - Review & Editing, Supervision, Funding acquisition.

Inconsistently consistent: Rationality is not reliable

Felix J. Nitsch, Luca M. Lüpken, Nils Lüschow, Tobias Kalenscher

Comparative Psychology, Institute of Experimental Psychology, Heinrich-Heine-University Düsseldorf

Author Note

Correspondence concerning this article should be addressed to Felix J. Nitsch. Email: felix.nitsch@hhu.de. The research was supported by a grant from the Deutsche Forschungsgemeinschaft (Bonn, Germany) to Tobias Kalenscher (grant-ID: DFG-KA 2675/4-3).

Abstract

Contemporarily, experimental investigations of revealed preference choice consistency utilize different tasks interchangeably. However, the reliability of choice consistency measurements among (inter-method) and within tasks (test-retest) has not been determined so far. Hence, it is unclear whether estimations of choice consistency fulfill a basic requirement of valid psychometric measures. Further, it is unclear how far results from different studies using different methodologies are comparable. In the study described here, we investigated the reliability of two established and one novel choice consistency tasks in an online-experiment under non-incentivized conditions in the choice domain of social decisions. Our results confidently indicate generally poor inter-method reliability and at best moderate test-retest reliability for the two indices, the Critical Cost Efficiency Index (CCEI) and the Houtman-Maks-Index (HMI), with the CCEI being the tentatively more reliable measure. This is especially concerning, since the full experiment (including test and retest measurement) lasted on average less than 45 minutes. Hence, it appears that estimations of choice consistency do not fulfill a basic requirement of valid psychometric measures. Further, results from different studies using different methodologies should not be compared without caution. Future work should investigate the impact of incentivization as well as the choice domain generality of our results.

Keywords: GARP, revealed preference, choice consistency, rationality, reliability

The Reliability of Choice Consistency

Revealed preference theory (Houthakker, 1950; Samuelson, 1938; Varian, 1982) assumes that any collection of choices by a rational decision maker can be reconciled with a definite structure of wants, cost efficiency and transitivity (Afriat, 1973). Specifically, for the typically considered two-dimensional choice objects (e.g. bundles of two different goods) the Generalized Axiom of Revealed Preference requires that if a decision maker accepts costs to choose a choice object over another (strict direct revealed preference), they may, *ceteris paribus,* in fact never choose the latter over the former choice object (no direct revealed preference) as long as it is not associated with higher costs (Banerjee & Murphy, 2006).

The most prominent way to quantify revealed preference choice consistency is Afriat's Critical Cost Efficiency Index (CCEI; Afriat, 1972; Varian, 1991): "The critical cost efficiency index (CCEI) is inspired by the fact that inconsistent choice behavior is not cost efficient. With given prices of choice options, the budget determines all affordable choice options. A revealed preference violation occurs when the decision maker does not choose the most preferred alternative that was affordable given the budget, but selects another less preferred option. This can be interpreted as a waste of money, as the decision maker did not obtain the maximum subjective value for [their] money. The critical cost efficiency index denotes the minimal hypothetical reduction of the budget of the decision maker necessary, so that all more preferred but not chosen options become unaffordable when the revealed preference violation occurs. A critical cost efficiency index of 1 denotes perfect consistency: The budget does not need to be reduced. The index approaches zero as the behavior becomes more inconsistent and the budget needs to be reduced starkly to eliminate inconsistency." (Nitsch & Kalenscher, 2020a). A drawback of the CCEI and conceptually similar indices is that it is entirely determined by the least cost-efficient choice and, thus, inherently is quite susceptible to outliers.

The Houtman-Maks-Index (HMI; Heufer & Hjertstrand, 2015; Houtman & Maks, 1985) on the other hand does not consider the severity of choice consistency violations but instead determines the size of the largest subset of choices consistent with GARP (or any other choice consistency axiom). Hence, the measure is possibly more robust to single outliers but also more sensitive to multiple but practically negligible violations.

Contemporarily, experimental investigations of revealed preference choice consistency utilize different tasks interchangeably. One line of research uses a task introduced by Choi et al. (2007). In their elaborate and widely used paradigm (henceforth “Diagram task”), participants must allocate a budget between two dimensions (e.g. two investment accounts, oneself and a co-player, apples and oranges) using a cartesian coordinate display. The task is mostly applied in the investigation of choices under risk (e.g Castillo et al., 2017; Cettolin et al., 2019; Choi et al., 2014; Kurtz-David et al., 2019) but also intertemporal choices (Chakraborty et al., 2017; Kim et al., 2018). It has the appeal that it transparently depicts all economic parameters (budget, prices, budget line etc.) and even allows for a visual identification of inconsistent choices. A potential drawback is that the task can be hard to understand for people without experience in the interpretation of diagrams and the theoretically large number of potential choice options.

Another line of research uses a more simplistic choice bundles task that was first prominently used by Harbaugh and colleagues (2001) and many others since (e.g. Bedi & Burghart, 2018; Burghart et al., 2013; Chung et al., 2017; Nitsch et al., 2021). In the choice bundles task (henceforth “Bundles task”), the budget line is divided into (equidistant) discrete points which are subsequently presented as a discrete set of choice options to the participant.

Conveniently, in this task participants can ignore the underlying economic parameters and must only choose the most liked choice bundle. This significantly reduces the cognitive demand of the task and is desirable for indivisible goods (e.g. food items), specific participant groups (e.g. children) and research questions (e.g. decisions under stress). A drawback of the task is that discrete choice options can only approximate optimal choices from a continuous budget line, which might introduce inconsistency by itself (see Dziewulski, 2018, for a psychophysical interpretation of choice consistency; see Nitsch & Kalenscher, 2020b for an application). A compromise (which so far has not been applied, however) would be to present participants with a slider (henceforth "Slider task"; see Methods), that allows for the continuous allocation of the budget while concealing economic parameters to a degree that allows for an intuitive approach to solving the task.

Problematically, however, the reliability of choice consistency measurements among (inter-method) and within tasks (test-retest) has not been determined so far. Hence, it is unclear how far estimations of choice consistency fulfill a basic requirement of valid psychometric measures. Further, it is unclear how far results from different studies using different methodologies are comparable. Lastly, it is an open question which type of consistency quantification, CCEI or HMI, is more desirable in terms of reliability. In the study described here, we investigated the reliability of social choice consistency measured via the three above described choice consistency tasks in an online-experiment under non-incentivized conditions. Specifically, we tested the following hypotheses:

H1: There is a significant, large correlation between the CCEI estimates of the Bundles, Slider and Diagram task. This means that the tasks are reliably measuring the same psychological construct.

H2: There is a significant, large test-retest reliability for CCEI estimates of the Bundles, Slider and Diagram task. This means that the tasks do reliably measure choice consistency for two subsequent measurements.

H3: There is a significant, large correlation between the HMI estimates of the Bundles, Slider and Diagram task. This means that the tasks are reliably measuring the same psychological construct.

H4: There is a significant, large test-retest reliability for HMI estimates of the Bundles, Slider and Diagram task. This means that the tasks do reliably measure choice consistency for two subsequent measurements.

H5: The CCEI and HMI show significantly different test-retest reliability.

Statistical and econometric power analyses confirmed sufficient power of our design. Analyses of the distribution of consistency indices showed that our participants, despite non-incentivization, behaved on average highly consistent (in line with previous reports in the literature) and significantly more consistent than bootstrapped random deciders. Given this, our results confidently indicate generally poor inter-method reliability and at best fair test-retest reliability for CCEI (H1 and H2) and HMI (H3 and H4), with the CCEI being the tentatively more reliable measure (H5). This is especially concerning, since the full experiment (including test and retest measurement) lasted on average less than 45 minutes. Hence, it appears that estimations of choice consistency do not fulfill a basic requirement of valid psychometric measures. Further, results from different studies using different methodologies should not be compared without caution.

Methods

Participants

101 adult, English-speaking participants completed our experiment. 48 were randomly assigned to an experimental manipulation group that is irrelevant to the presented research question and were, thus, discarded for all analyses. No other participants were excluded resulting in a final sample size of N = 53 participants. Table 1 gives an overview of the demographics.

Procedure & Design

Participants were recruited via Prolific receiving a compensation of 4.50 pounds. The experiment was programmed in jsPsych (de Leeuw & Motz, 2016) and hosted on Pavlovia. Before the start of the experiment, all participants were fully debriefed about the content and aim of the research project and provided informed consent via a check box. After providing consent, we asked for their demographic information. Next, participants underwent the first measurement of all three experimental tasks in randomized order. For the first measurement, each task entailed a detailed description and 5 practice trials. After completion of the first measurement participants solved a filler task that consisted of reading three informational texts about unrelated topics and answering three quiz questions on the content of the texts (see Appendix B). Then participants underwent the second measurement of all three experimental tasks, again in randomized order. At the end of the experiment, participants answered several questions regarding their decision strategies and experiences solving the tasks. Then they were redirected back to Prolific to receive their compensation.

Our experimental design was completely within-subject. Participants solved all three decision tasks for two measurements (3 x 2 within-subject design).

Experimental Tasks

Generally, all decision tasks consisted of $I = 20$ decisions per measurement where participants, hypothetically, had to allocate a budget m_i between them and their best friend, resulting in a final monetary split of $x_i = (x_i^{Self}, x_i^{Friend})$. Importantly, the monetary endowment m_i and the "prices" of keeping and giving money $p_i = (p_i^{Self}, p_i^{Friend})$varied per decision. Hence, $x_i^{Self} = \frac{share_i^{Self} m_i}{p_i^{Self}}$and $x_i^{Friend} = \frac{share_i^{Friend} * m_i}{p_i^{Friend}}$, with the share indicating the relative fraction of the budget (0 to 1) allocated to each account. Budgets and prices were randomly sampled per trial: $m_i \in [2, 3, 4, 5, 6, 7, 8, 9, 10]$ and $p_i^{Self}, p_i^{Friend} \in [1, 2, 3]$. Note, that for our analysis we normalized prices and budgets so that $\sum p_i = 1$and $m_i = x_i^{Self} p_i^{Self} + x_i^{Friend} p_i^{Friend}$.

For each task and measurement, we further included two attention check trials where participants were instructed to allocate the full budget either to themselves or their best friend. Those trials were not included in the analysis. If participants failed an attention check for a given measurement of a task, we excluded that measurement of the task from our analysis specifically.

A screenshot of a trial from each task is available in Appendix A.

Diagram Task

For each decision participants had to choose a point on a diagonal line in a coordinate system. The points on the diagonal line represented the possible money allocations between them and their best friend that they might choose. In each coordinate system, the vertical axis corresponded to the money chosen for themselves ("You"), and the horizontal axis corresponded to the money chosen for their best friend ("Friend"). While they were making their decision, they could see which amount of money they had chosen for themselves and for their best friend in the upper right corner of the coordinate system. The flatter the lines, the more money their best friend could receive as a maximum compared to them. The steeper the lines are, the more money they could receive as a maximum compared to their best friend.

Bundles Task

For each decision participants had a choice of 5 different money allocations and were instructed to simply choose the allocation that they thought was best.

Slider Task

For each decision participants had to choose a point on a horizontal slider, which represented the possible allocations of money amounts between them and their best friend. While making their decision, they could see which amount of money they had chosen for themselves and their best friend in two boxes above the slider. The labeling of the endpoints and spatial presentation was randomized from round to round.

Task Questionnaires

At the end of the experiment participants were asked to answer how they reached their decisions ("How did you reach your decisions?") and what they considered particularly important in their decisions ("What was particularly important to you in your decisions?") in open text format. Further they were asked multiple questions regarding their experiences with the specific task formats that will be reported elsewhere.

Analysis

Revealed Preference Analysis

Let N be the number of different commodity types in a commodity bundle. Let X be the non-negative, N-dimensional space of commodity bundles. Let P be the non-negative, N -dimensional space of prices of commodities. Let M be the non-negative, one-dimensional space of budgets. Let $I = i, j, \ldots, n$. denote observations of choice. Let x_i be the chosen commodity bundle of an observation i. Each bundle x_i is a N-dimensional vector of the shape $x_i = (x_i^1, x_i^2, \ldots, x_i^n)$, with each scalar component x_i^n representing the quantity of commodity type n within bundle x_i. Let p_i be the given prices of commodities of an observation i. Each prices p are a *N*-dimensional vector of the shape $p_i = (p_i^1, p_i^2, \ldots, p_i^n)$,, with each scalar component p_i^n representing the price of commodity type n per unit size. Then the scalar product $x_i \bullet p_i$ represents the total price of a commodity bundle x_i at some prices p_i. Let m_i be the given budget of an observation i. We assume, that a decision maker spends all their budget so that $x_i \bullet p_i = m_i$.

Definition 1 (Direct Revealed Preference). A bundle x_i is directly revealed preferred to another bundle x_j if and only if $x_j \bullet p_i \leq m_i$. Then we denote $x_i R_D x_j$.

Definition 2 (Revealed Preference). A bundle x_i is revealed preferred to another bundle x_k if there exists a transitive preference relation $x_i R_D x_j R_D x_k$ between both bundles.We denote $x_i R x_k$.

Definition 3 (Strict Direct Revealed Preference). A bundle x_i is strictly directly revealed preferred to another bundle x_j if and only if $x_j \bullet p_i < m_i$. Then we denote $x_i P_D x_j$.

Axiom 1 (Generalized Axiom of Revealed Preference). $x_i R x_j \Leftrightarrow \neg x_j P_D x_i$.

In this framework, the CCEI presents a relaxation of Defintion 3, so that only $x_i P_D x_j \leftarrow x_j \bullet p_i < CCEI * m_i$ is required. The CCEI is then the highest possible value so that

Axiom 1 holds for all observations. The HMI on the other hand denotes the largest number of observations $HMI \leq I$for which Axiom 1 holds.

Statistical Analysis

To quantify the reliability of choice consistency we calculated the Pearson correlation coefficient between the variables of interest. Reliability indicates how much of the total variance in the variable of interest is not caused by measurement error. A reliability of 1 indicates a perfect measurement: all variance in the measurement is caused by true differences in the variable of interest. As reliability approaches zero, the measurement is becoming less precise and a higher fraction of variance is due to measurement error (Kimberlin & Winterstein, 2008). In accordance with our hypotheses, we tested for a positive correlation using one-sided t-tests. Specifically, for hypothesis 1 and 3 we calculated the correlation of consistency measures among tasks for each measurement (3 x 2 = 6 comparisons). For hypothesis 2 and 4 we calculated the correlation of consistency measures within tasks across measurement (3 comparisons). To test whether the HMI is a more reliable measure than the CCEI (hypothesis 5), we pairwisely tested for differences in the test-retest reliability of CCEI compared to HMI per task (3 comparisons; Diedenhofen & Musch, 2015). The level of statistical significance was set *a priori* to alpha = .05. However, since we conducted multiple comparisons for each hypothesis, we applied a Bonferroni correction resulting in a significance-threshold of alpha = .05/6 =~ .008 for hypotheses 1 and 3, and alpha = .05/3 =~ .017 for hypotheses 2, 4 and 5. Further, for qualitative interpretation of reliability we will adhere to the following standards: a reliability coefficient of below .40 is considered poor; when it is between .40 and .59 it can be considered fair; when it is between .60 and .74, it can be considered good; and when it is between .75 and 1.00 it can be considered excellent (Cicchetti, 1994).

Power Analysis

We *a priori* defined the required sample size of our experiment using a statistical power analysis in G*Power (Erdfelder et al., 1996). As mentioned above, we would only deem a correlation between and within tasks large enough, if the effect size is r >= 0.5. Additionally, we aimed for a statistical power of 1-Beta = .90 and considered the lowest Bonferroni-corrected significance level of alpha = .05/6 =~ .008. Results indicated that we would need at least 44 participants to reach our power goal.

Bronars' Power

To determine the statistical power of our GARP test we bootstrapped 1000 virtual participants from our dataset (Bronars, 1987). Results showed that Bronars Power = 91,8% bootstrapped participants did not pass GARP.

Qualitative Content Analysis

In order to gain a further understanding of the decision-making process and to further validate our data we conducted an inductive, qualitative content analysis (Mayring, 2004) using the two free-text responses about the decision strategy of our participants. For our analysis, we concatenated the answers of our participants to both questions. Using an inductive approach, 5 exhaustive and mutually-exclusive categories were generated. Lastly, we compared the relative frequencies of categories to identify key decision-making strategies.

Results

Reality Check

To ensure that our collected choice data can be meaningfully interpreted we tested for a difference in choice consistency of our participants to a benchmark of 53 bootstrapped participants using a non-parametric Mann-Whitney-U test. Results indicated that our participants were significantly more consistent in their choices for all tasks and measurements according to both CCEI and HMI (see table 2 for descriptive statistics and significance tests, and figure 1 and figure 2 for the full distribution of the data).

H1: Inter-method Reliability of CCEI

Our results indicate poor inter-method reliability, with 1 out of 6 comparisons showing a significant correlation of the measured CCEI values (see table 3) and a mean correlation of r=0.237.

H2: Test-Retest Reliability of CCEI

2 out of 3 comparisons showed a significant correlation of the measured CCEI values (see table 3). Further, results indicated a mean correlation of r=0.459 (r=0.537 excluding the novel Slider task), which can be considered fair. The average absolute difference between first and second measurement amounted to M_delta_CCEI = 0.084 for the Diagram task, M_delta_CCEI = 0.111 for the Bundles task and M_delta_CCEI = 0.113 for the Slider task (see Appendix C for visualization).

H3: Inter-method Reliability of HMI

Our results indicate poor inter-method reliability, with 1 out of 6 comparisons showing a significant correlation of the measured CCEI values (see table 4) and a mean correlation of r=0.254.

H4: Test-Retest Reliability of HMI

1 out of 3 comparisons showed a significant correlation of the measured HMI values (see table 4). Further, results indicated a mean correlation of r=0.428 (r=0.471 excluding the novel Slider task), which can be considered fair. The average absolute difference between first and

second measurement amounted to M_delta_HMI = 0.791 for the Diagram task, M_delta_HMI = 1.042 for the Bundles task and M_delta_CCEI = 0.837 for the Slider task (see Appendix C for visualization).

H5: Test-Retest Reliability Comparison of CCEI and HMI

Our results indicate a significantly higher test-retest reliability of the CCEI compared to the HMI for the Diagram task (z =-2.283, p=.022), but neither for the Bundles task (z =0.858, p=.391) nor the Slider task (z=0.502, p=.616).

Qualitative content analysis of reported decision strategies

Our results indicated that the vast majority of participants either tried to fairly share the payout (23 participants, 43.396%) or maximize the total payout (20 participants, 37.736%). Few participants decided with an egotistical (6 participants, 11.321%) or prosocial bias (2 participants, 3.774%). For 2 participants (3.774%) we could not determine a strategy from their response. For an overview see figure 3.

Discussion

So far the reliability of revealed preference choice consistency measurements among (inter-method) and within tasks (test-retest) has remained an untouched topic. Any interpretations of rationality as a valid psychometric trait variable, as well as comparisons between studies of different methodology, therefore, relied on a leap of faith in this aspect. In the study described here, we investigated the reliability of social choice consistency measured via three choice consistency tasks in an online-experiment under non-incentivized conditions. We operationalized choice consistency via two popular and conceptually dissimilar quantitative indices, namely Afriat's critical cost efficiency index (CCEI) and the Houtman-Maks-Index (HMI).

Statistical and econometric power analysis confirmed sufficient power of our design. Analyses of the distribution of consistency indices showed that our participants, despite non-incentivization, behaved on average highly consistent and significantly more consistent than bootstrapped random deciders. Given this, our results confidently indicated generally poor inter-method reliability and at best fair test-retest reliability for CCEI (H1 and H2) and HMI (H3 and H4), with the CCEI being the tentatively more reliable measure (H5). This is especially concerning, since the full experiment (including test and retest measurement) lasted on average less than 45 minutes. Hence, it appears that estimations of choice consistency do not fulfill a basic requirement of valid psychometric measures. Further, results from different studies using different methodologies should not be compared without caution. However, there are three important limitations to our study.

First and foremost, neither of our choice tasks was incentive-compatible, which is an important feature in economic decision-making. While we cannot rule out that our participants might have behaved differently under full incentivization, we can robustly show that our participants behaved significantly more consistently than random bootstrapped deciders. Further, participants' introspections on their decision strategies do not raise concerns regarding artificial or non-interpretable response behaviour (see figure 3). Lastly, we imposed higher standards of valid choices than previous studies by implementing multiple attention checks. Therefore, overall, our data does not provide evidence for random or artificial behaviour.

The second limitation is that we only used tasks in the social domain of choice, similar to Andreoni et al. (2002). Hence, we cannot speak to the domain generality of our results, which is an important question for future research as we recently formulated elsewhere (Nitsch & Kalenscher, 2020a).

The third limitation is that we only investigated reliability for two out of many (however, conceptually closely related) consistency indices, e.g. the Money Pump Index (Echenique et al., 2011), Varian's Index (Varian, 1991) or the Minimum Cost Index (Dean & Martin, 2016). While an exhaustive evaluation of all existent consistency indices is beyond the scope of this paper, our openly shared dataset (https://doi.org/10.17605/OSF.IO/QNFTU) allows for the conceptual replication of our results via other consistency measures.

In conclusion, despite the above mentioned limitations, our results indicate that rationality in the interpretation of revealed preference consistency does not meet a fundamental requirement of valid psychometric measures, i.e. reliability. This suggests that choice consistency might rather be considered a characteristic of a specific collection of choices than an underlying trait variable of the decision-maker. This calls for caution in the interpretation of previous results in the field.

CRediT author statement

Felix J. Nitsch: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Visualization, Supervision, Project Administration. **Luca M. Lüpken:** Conceptualization, Software, Investigation, Writing - Review & Editing. **Nils Lüschow:** Software, Writing - Review & Editing. **Tobias Kalenscher:** Conceptualization, Writing - Review & Editing, Supervision, Funding acquisition.

Open Practices

Anonymized raw and processed data and analysis scripts are available at: https://doi.org/10.17605/OSF.IO/QNFTU. The experiment was not pre-registered.

References

Afriat, S. N. (1972). Efficiency Estimation of Production Functions. *International Economic Review*, *13*(3), 568–598. JSTOR. https://doi.org/10.2307/2525845

Afriat, S. N. (1973). On a system of inequalities in demand analysis: An extension of the classical method. *International Economic Review*, 460–472. https://doi.org/10.2307/2525934

Andreoni, J., & Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica*, *70*(2), 737–753. https://doi.org/10.1111/1468-0262.00302

Banerjee, S., & Murphy, J. H. (2006). A simplified test for preference rationality of two-commodity choice. *Experimental Economics*, *9*(1), 67–75. https://doi.org/10.1007/s10683-006-4313-6

Bedi, G., & Burghart, D. R. (2018). Is utility maximization compromised by acute intoxication with THC or MDMA? *Economics Letters*, *171*, 128–132. https://doi.org/10.1016/j.econlet.2018.06.021

Bronars, S. G. (1987). The power of nonparametric tests of preference maximization. *Econometrica: Journal of the Econometric Society*, 693–698. https://doi.org/10.2307/1913608

Burghart, D. R., Glimcher, P. W., & Lazzaro, S. C. (2013). An expected utility maximizer walks into a bar... *Journal of Risk and Uncertainty*, *46*(3), 215–246. https://doi.org/10.1007/s11166-013-9167-7

Castillo, M., Dickinson, D. L., & Petrie, R. (2017). Sleepiness, choice consistency, and risk preferences. *Theory and Decision*, *82*(1), 41–73.

https://doi.org/10.1007/s11238-016-9559-7

Cettolin, E., Dalton, P. S., Kop, W. J., & Zhang, W. (2019). Cortisol meets GARP: The effect of stress on economic rationality. *Experimental Economics*. https://doi.org/10.1007/s10683-019-09624-z

Chakraborty, A., Calford, E. M., Fenig, G., & Halevy, Y. (2017). External and internal consistency of choices made in convex time budgets. *Experimental Economics*, *20*(3), 687–706. https://doi.org/10.1007/s10683-016-9506-z

Choi, S., Fisman, R., Gale, D., & Kariv, S. (2007). Consistency and heterogeneity of individual behavior under uncertainty. *American Economic Review*, *97*(5), 1921–1938. https://doi.org/10.1257/aer.97.5.1921

Choi, S., Kariv, S., Müller, W., & Silverman, D. (2014). Who Is (More) Rational? *American Economic Review*, *104*(6), 1518–1550. https://doi.org/10.1257/aer.104.6.1518

Chung, H.-K., Tymula, A., & Glimcher, P. (2017). The Reduction of Ventrolateral Prefrontal Cortex Grey Matter Volume Correlates with Loss of Economic Rationality in Aging. *The Journal of Neuroscience*, *37*(49), 1171–17. https://doi.org/10.1523/JNEUROSCI.1171-17.2017

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284.

de Leeuw, J. R., & Motz, B. A. (2016). Psychophysics in a Web browser? Comparing response times collected with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research Methods*, *48*(1), 1–12. https://doi.org/10.3758/s13428-015-0567-2

Dean, M., & Martin, D. (2016). Measuring rationality with the minimum cost of revealed preference violations. *Review of Economics and Statistics*, *98*(3), 524–534.

https://doi.org/10.1162/REST_a_00542

Diedenhofen, B., & Musch, J. (2015). cocor: A Comprehensive Solution for the Statistical Comparison of Correlations. *PLOS ONE*, *10*(4), e0121945. https://doi.org/10.1371/journal.pone.0121945

Dziewulski, P. (2018). *Just-noticeable difference as a behavioural foundation of the critical cost-efficiency index*. https://www.economics.ox.ac.uk/materials/working_papers/4603/848-dziewulski.pdf

Echenique, F., Lee, S., & Shum, M. (2011). The money pump as a measure of revealed preference violations. *Journal of Political Economy*, *119*(6), 1201–1223. https://doi.org/10.1086/665011

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 1–11.

Harbaugh, W. T., Krause, K., & Berry, T. R. (2001). GARP for kids: On the development of rational choice behavior. *American Economic Review*, *91*(5), 1539–1545. https://doi.org/10.1257/aer.91.5.1539

Heufer, J., & Hjertstrand, P. (2015). Consistent subsets: Computationally feasible methods to compute the Houtman–Maks-index. *Economics Letters*, *128*, 87–89. https://doi.org/10.1016/j.econlet.2015.01.024

Houthakker, H. S. (1950). Revealed preference and the utility function. *Economica*, *17*(66), 159–174. https://doi.org/10.2307/2549382

Houtman, M., & Maks, J. (1985). Determining all maximal data subsets consistent with revealed preference. *Kwantitatieve Methoden*, *19*(1), 89–104.

Kim, H. B., Choi, S., Kim, B., & Pop-Eleches, C. (2018). The role of education interventions in improving economic rationality. *Science*, *362*(6410), 83–86. https://doi.org/10.1126/science.aar6987

Kimberlin, C. L., & Winterstein, A. G. (2008). Validity and reliability of measurement instruments used in research. *American Journal of Health-System Pharmacy*, *65*(23), 2276–2284. https://doi.org/10.2146/ajhp070364

Kurtz-David, V., Persitz, D., Webb, R., & Levy, D. J. (2019). The neural computation of inconsistent choice behavior. *Nature Communications*, *10*(1), 1–14. https://doi.org/10.1038/s41467-019-09343-2

Nitsch, F. J., & Kalenscher, T. (2020a). *Keeping a cool head at all times. What determines choice consistency?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/etyhx

Nitsch, F. J., & Kalenscher, T. (2020b). *Influence of memory processes on choice consistency.* https://doi.org/10.17605/OSF.IO/AKGT9

Nitsch, F. J., Sellitto, M., & Kalenscher, T. (2021). The effects of acute and chronic stress on choice consistency. *Psychoneuroendocrinology*, *131*, 105289. https://doi.org/10.1016/j.psyneuen.2021.105289

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71.

Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, 945–973. https://doi.org/10.2307/1912771

Varian, H. R. (1991). *Goodness-of-fit for revealed preference tests*. Department of Economics, University of Michigan Ann Arbor.

Tables

Table 1

Demographics

Variable	Frequency	Percent
Total	53	
Gender		
Male	23	56.604%
Female	20	43.396%
Education		
No high school diploma	1	1.887%
High school diploma	22	41.509%
Jr./Comm. College degree / Undergraduate degree	23	43.396%
Masters Degree / Professional schools degree (Law, Medicine, etc.)	7	13.208%
Monthly Net Income (in Pounds)		
0-499	16	30.189%
500-999	13	24.528%
1000-1499	8	15.094%
1500-1999	3	5.660%
2000-2999	5	9.434%
>3000	3	5.660%
Omitted	5	9.434%

Table 2

Choice consistency of participants compared to bootstrapped virtual deciders

Task 1	Measurement	Index	M (SD)	W	p
Diagram	1	CCEI	0.904 (0.183)	2309	<.001*
Bundles	1	CCEI	0.858 (0.157)	2162	<.001*
Slider	1	CEEI	0.891 (0.166)	2289	<.001*
Diagram	2	CCEI	0.929 (0.169)	2457	<.001*
Bundles	2	CCEI	0.867 (0.151)	2135	<.001*
Slider	2	CEEI	0.935 (0.148)	2488	<.001*
Diagram	1	HMI	19.065 (1.357)	2366	<.001*
Bundles	1	HMI	18.388 (1.483)	1979	<.001*
Slider	1	HMI	19.250 (1.014)	2476.5	<.001*
Diagram	2	HMI	19.458 (0.922)	2574	<.001*
Bundles	2	HMI	18.692 (1.322)	2098.5	<.001*
Slider	2	HMI	19.365 (1.172)	2504	<.001*
Bootstrap		CCEI	0.738 (0.151)		
Bootstrap		HMI	17.906 (1.197)		

Note: The asterisk denotes significance at p<.05/12=.004 (Bonferroni corrected for 12 parallel tests).

Table 3

Reliability of the Critical Cost Efficiency Index

Task 1	Task 2	Measurement	t(df)	r	p
Diagram	Bundles	1	t(40)=0.043	0.007	.483
Diagram	Slider	1	t(37)=1.250	0.201	.110
Bundles	Slider	1	t(39)=-0.284	-0.045	.611
Diagram	Bundles	2	t(46)=0.752	0.110	.228
Diagram	Slider	2	t(45)=7.866	0.761	<.001*
Bundles	Slider	2	t(49)=1.248	0.176	.109
Diagram	Diagram	1 & 2	t(41)=5.088	0.622	<.001*
Bundles	Bundles	1 & 2	t(46)=3.324	0.440	<.001*
Slider	Slider	1 & 2	t(41)=1.864	0.280	.035

Note: The asterisk denotes significance according to the predefined threshold per hypothesis.

Table 4

Reliability of the Houtman-Maks-Index

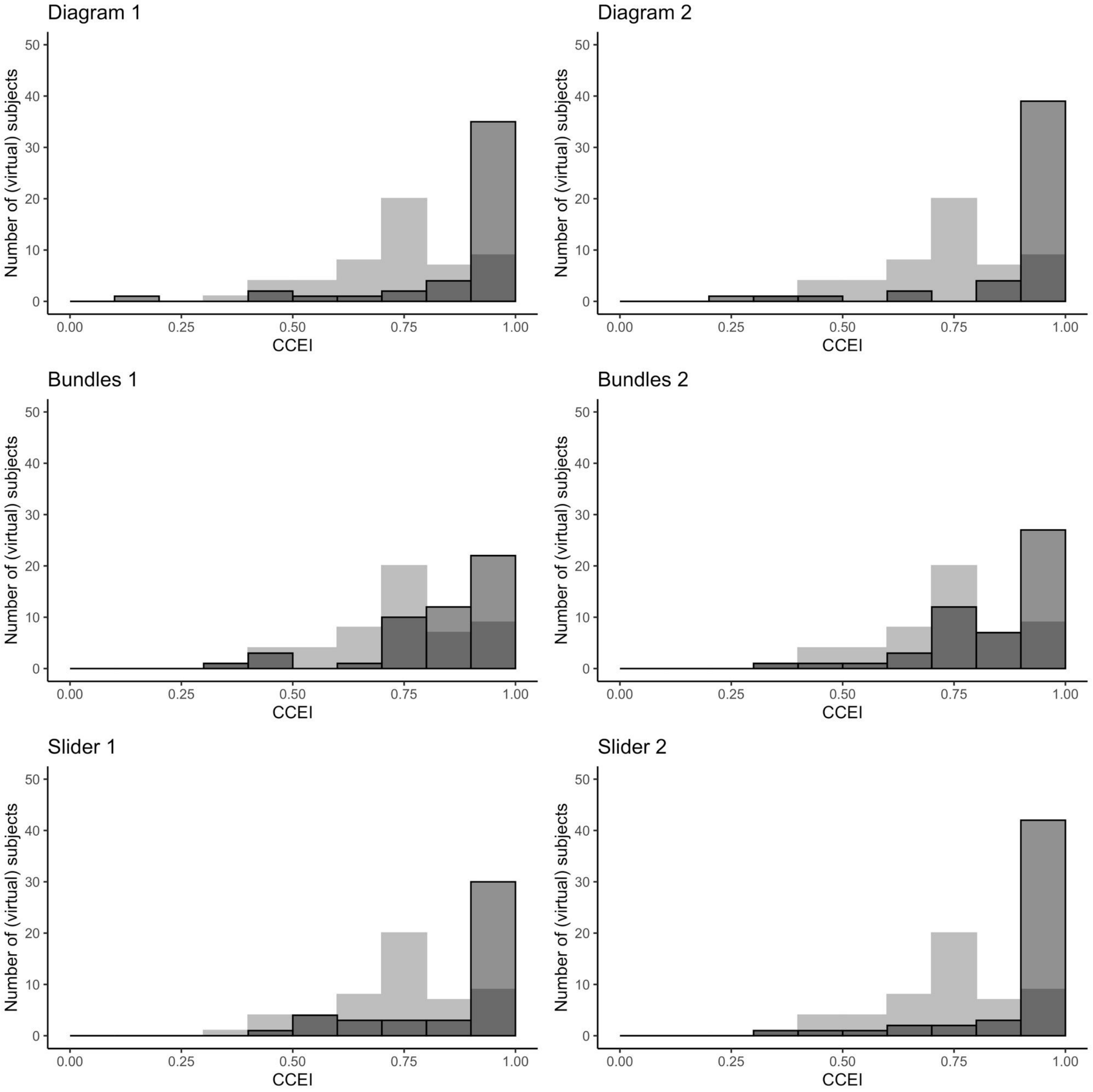
Task 1	Task 2	Measurement	t(df)	r	p
Diagram	Bundles	1	t(40)=0.438	0.069	.332
Diagram	Slider	1	t(37)=0.690	0.113	.247
Bundles	Slider	1	t(39)=0.422	0.067	.338
Diagram	Bundles	2	t(46)=1.577	0.226	.061
Diagram	Slider	2	t(45)=7.584	0.749	<.001*
Bundles	Slider	2	t(49)=0.743	0.106	.230
Diagram	Diagram	1 & 2	t(41)=2.651	0.383	0.006*
Bundles	Bundles	1 & 2	t(46)=4.468	0.550	<.001*
Slider	Slider	1 & 2	t(41)=2.296	0.338	.013*

Note: The asterisk denotes significance according to the predefined threshold per hypothesis.

Figures

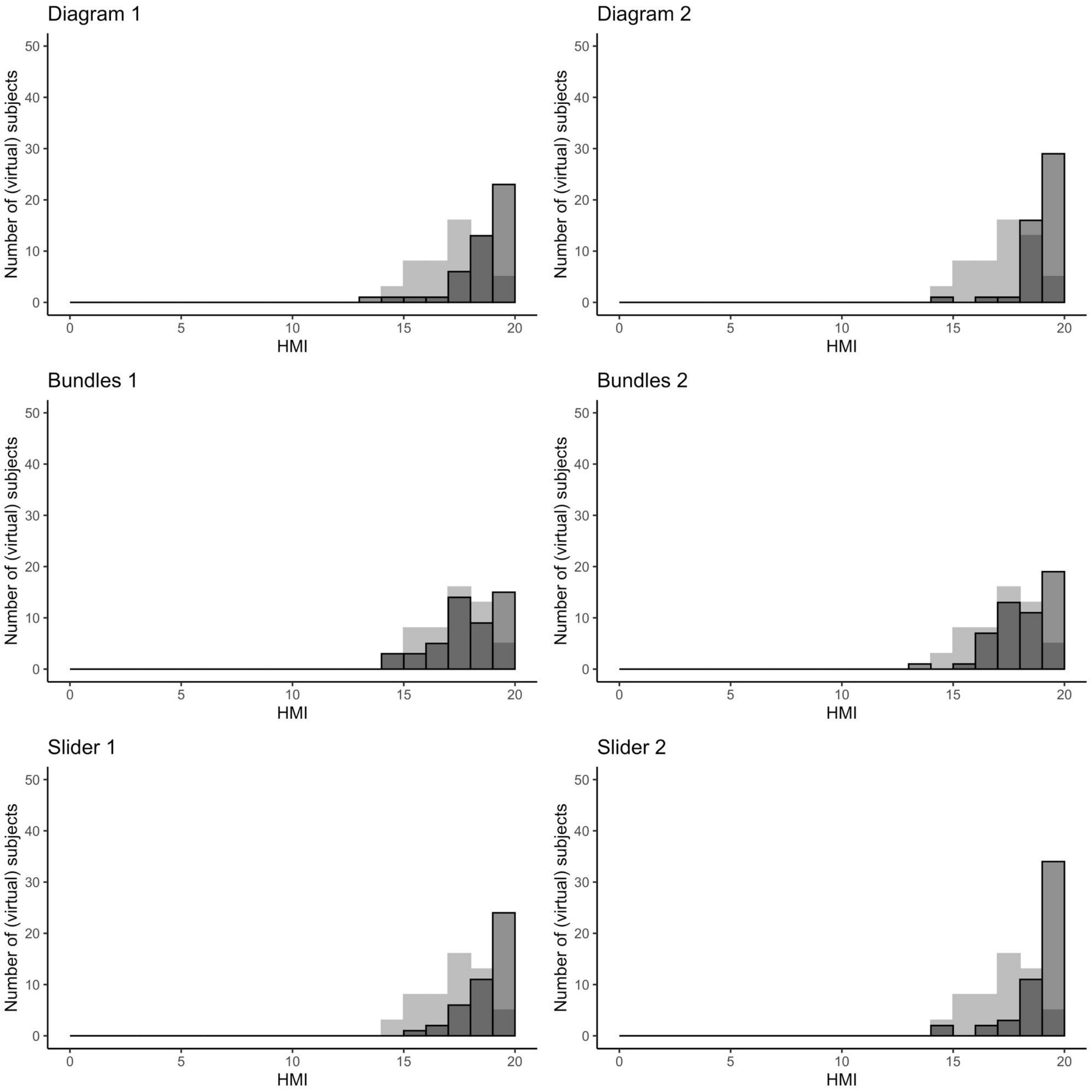
Figure 1

Distribution of CCEI in participants and bootstrapped virtual participants



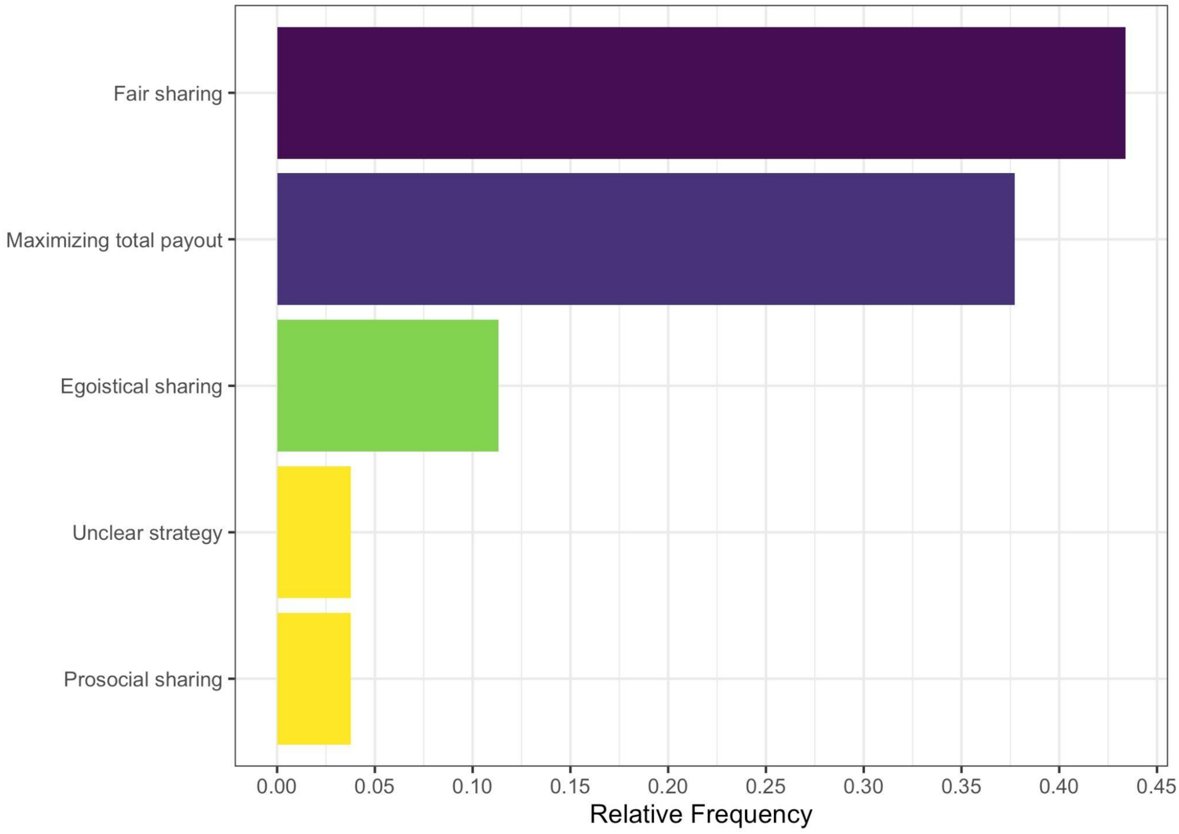
Histograms of the Critical Cost Efficiency Index (CCEI) distribution of our participants (dark grey) in comparison to bootstrapped random deciders (light grey) for the 3 different tasks and 2 measurements. In all cases, real participants behaved significantly more consistently than the bootstrapped random deciders. Furthermore, the absolute levels of consistency are in alignment with consistency levels reported in the literature.

Figure 2
Distribution of HMI in participants and bootstrapped virtual participants



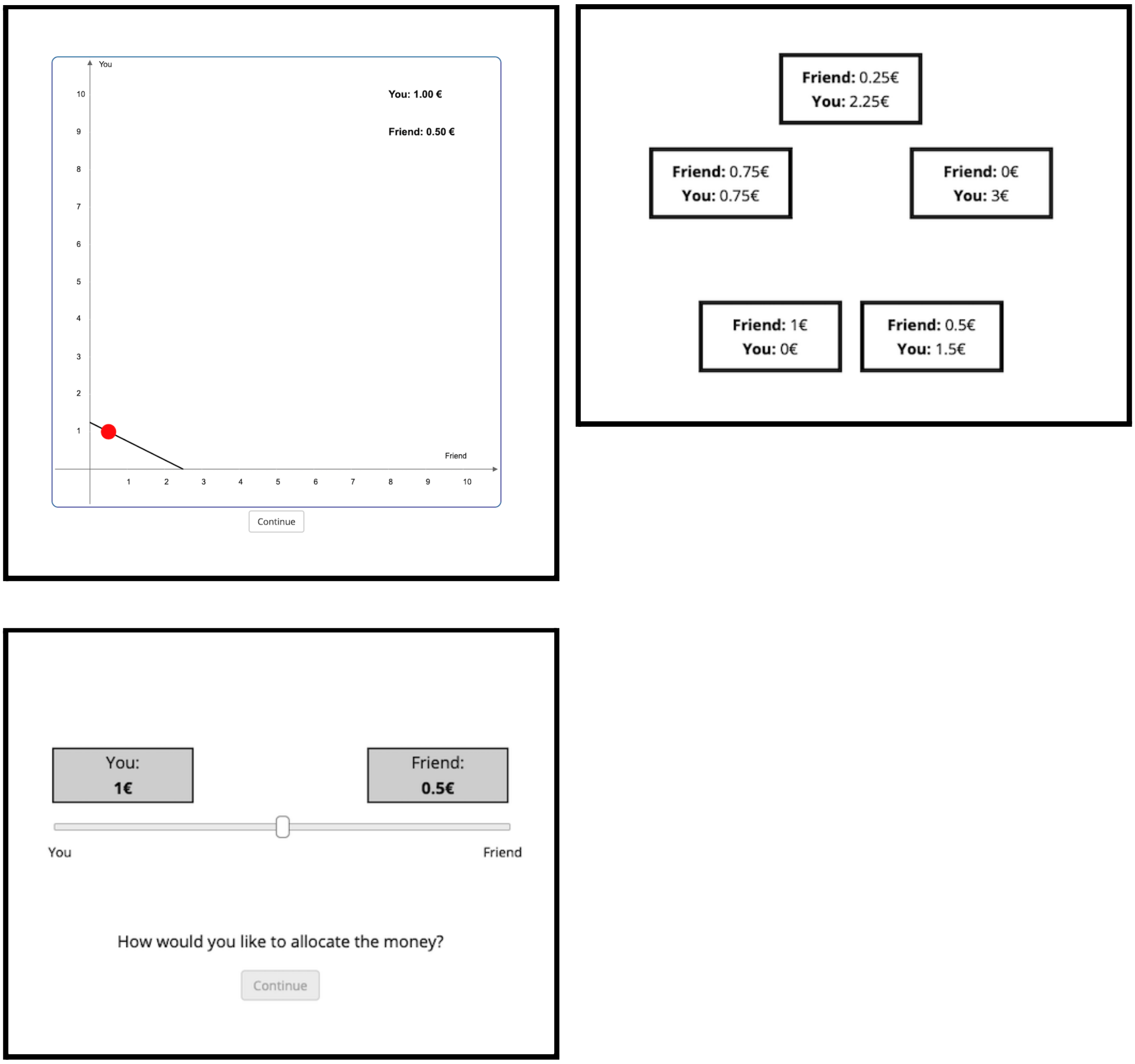
Histograms of the Houtman-Maks-Index (HMI) distribution of our participants (dark grey) in comparison to bootstrapped random deciders (light grey) for the 3 different tasks and 2 measurements. In all cases, real participants behaved significantly more consistently than the bootstrapped random deciders. Furthermore, the absolute levels of consistency are in alignment with consistency levels reported in the literature.

Figure 3
Qualitative content analysis of self-reported decision strategies



Relative frequency of self-reported decision strategy categories (N = 53). Results do not raise concerns regarding artificial or non-interpretable response behaviour.

Appendix A - Screenshots from Diagram, Bundles and Slider task



You: 1.00 €

Friend: 0.50 €

Friend: 0.25€
You: 2.25€

Friend: 0.75€
You: 0.75€

Friend: 0€
You: 3€

Friend: 1€
You: 0€

Friend: 0.5€
You: 1.5€

You:
1€

Friend:
0.5€

How would you like to allocate the money?

Depicted are the Diagram task (upper left), the Bundles task (upper right) and the Slider task (lower left).

Appendix B - Filler Task

Text 1: I start in Munich. I want to go to the Gassmanns. Alexandra Gasmann was nice on the phone, but also direct. She has nine children and she's a CDU (German political party) woman. She will tell me off, just think of Seehofer with Röttgen (German politicians) back then. They know how to do that. I'm here because sometimes I would like to have an even bigger family, but my wife doesn't really, and our life is also not exactly what I would call zen. I will test this, being a "dad for a day", because the real dad is at work until noon. Let's do it! Mrs. Gasmann gets up at 5 am every day because she also just wants to take a breath for five minutes in the morning without hearing "Mommy" right away and having to be there. It's her chill area in the kitchen while making sandwiches from 5-6 am in the morning. I get that. Samantha has to go to school, Ferdinand too, Jakob is still hanging around before daycare starts, and I'm about to take Gwendoline to the school bus. The older ones are already on their way. I'll say it right away, this is not a posh family, wrong channel. I have to drag the toddler to daycare, he doesn't want to walk at all. His peers and his friends don't have as many siblings as he does, but he thinks it's "cool", and he wouldn't want to be alone with mom and dad under any circumstances. Next, I go grocery shopping with Mrs. Gasmann and ask her if she sometimes gets stupid remarks. "Don't you have a TV at home? Did you really need that now, too?". Remarks like "Are you too stupid to use contraception" are part of the daily routine. Mrs. Gasmann has days on which such remarks bounce off of her and days on which she gets really grumpy. However, when asked if she wants to have more children, she replies that she would be very happy if that worked out. Alexandra buys huge bags of groceries every day. How do they manage with the money? They have a fantastic rental contract, then child support, Alexandra is paid a little for her work in the district parliament and her husband Arthur is a master butcher. It is enough for private schools, but food is mostly from the discounter. Alexandra buys 22 pounds of potatoes and about 125-134 cups of milk a week. The Gasmanns have two apartments: Downstairs the youngest live with the parents, upstairs the older kids, and there is also the laundry room. They have three washing machines and Alexandra washes 18 loads a week and also irons everything. She always has the best ideas, she tells me.

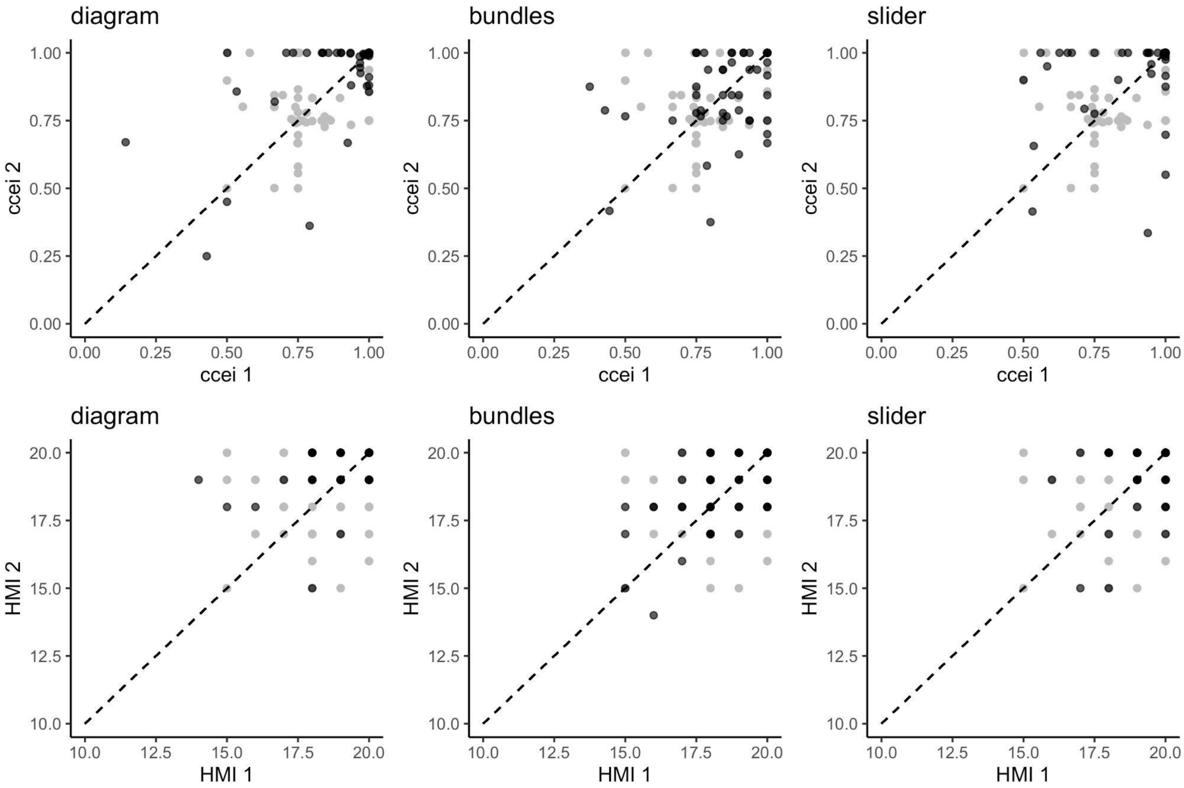
Text 2: "You are what you eat", goes the popular saying. It means what we eat is the cornerstone of our life. It determines our health, our mobility and also our education. But you can't teach an old dog new tricks. In other words, if you don't learn to eat healthy as a child, you won't learn later, and then you won't teach your children either. In Germany it is particularly serious. There is hardly any other European country with so many malnourished adults and children. Fortunately, there is school, where you learn for life. In the best case scenario, nutrition would be a compulsory subject and the students would go to the cafeteria with their teachers after class. There, the food would be prepared by trained chefs, and the students could immediately apply what they have just learned. But the reality is different. In Germany, it has to be practical - and cheap. More than a third of the schools we surveyed complained about a budget that was too tight. On average, a meal may only cost students €2.50, but a healthy meal costs at least €4 per serving. In many schools, good quality is, therefore, not possible. Most of the time, the food is cooked elsewhere and then delivered hot, so it is kept warm for several hours. This is also the case in about half of the schools we surveyed, including two-thirds of the elementary schools. What finally ends up on the plate is overcooked with no trace of vitamins. There is hardly any variety. If several dishes are offered at all, the students often have to decide a week in advance what they would like to eat. By far not every school, where students eat, has its own cafeteria, and if one exists, it is often cold, cramped or uncomfortable. Churches are no different, to save money, the food is often prepared and served by unlearned and poorly paid helpers. Hygiene regulations are rarely known, or adequate washing or sanitizing facilities are lacking. As a result, no one likes to go there. If they can, students prefer to go to the kiosk instead and eat there. According to experts, it would cost the state about 500 million euros a year to provide all students with a healthy lunch. That is only a fraction of what is spent on military weaponry every year. Why are investments made for this, but not for the children? In fact, responsibility is shifted from the state to the communities and often from the communities to the schools. They, in turn, transfer much of the responsibility to caterers who are supposed to deliver cheaply. No one wants to be responsible. Yet, good school catering would have many advantages. The children would not only be healthier and perform better, but they would also know more about healthy eating. In the long run, this would even save money, because, in the future, there would be fewer cases of illness society would have to pay for, and there would be more educated people who also earn more. The bottom line is a healthier and richer society.

Text 3: This is me, Tobias. Normally I see the world from 6 feet. For seven days, I want to leave this perspective and explore the world of the little ones. This daycare here in Otterndorf on the North Sea is perfect for that. 130 children come and go here. What do they do all day? What does life look like from a height of 3'6''? What do children think of our adult world? What can I perhaps learn from them? - That's what this text is about. Food determines the time reckoning here. There are three times of day: Before the meal, after the meal and of course the meal itself. The entire process is accompanied by an impressive ritual: "Beep, beep, little mouse, come out of your little house. Little mouse says "beep, bon appétit!". On my first day, the little ones eat large amounts of rice and a strange brown sausage gravy. The rest of the week, the menu includes woodruff jello and whole wheat pasta with custard. All that high-energy food must of course be broken down, preferably by “fighting”. Imagination is the best recipe. Everyone used to have it. The children say they have coffee here, but that it doesn't taste good and that you get a stomach ache if you drink too much of it. Rationality always comes at the expense of fun. This sobering principle becomes clear when adult voices interfere with the little ones' plans and routines: The kids have to help clean the bathroom if they don't behave well. During sports, it quickly becomes evident that with the first leap over the box, the disappointment of earlier is forgotten: Life is always lived in the here and now. The moment counts. All the things that we adults have to painstakingly relearn in yoga-zen-mindfulness seminars, they can all do here, just like that. When did I lose that? A daycare day lasts about three adult days, but at some point, it is still closing time, usually around 5 pm.

Memory test:

1. What time does Mrs. Gassmann get up every morning?
    1. 5 am
    2. 6 am
    3. 7 am
2. How much would it cost the German government to provide a healthy lunch?
    1. 200 million pounds
    2. 300 million pounds
    3. 500 million pounds
3. According to the children, which drink causes a stomach ache?
    1. Coffee
    2. Coke
    3. Beer

Appendix C - Test-Retest Reliability



Depicted are the test-retest reliability of each task for the CCEI and HMI (in black). For comparison, in grey, 53 bootstrapped deciders are depicted. The dashed line indicates optimal test-retest reliability (r=1).

General Discussion

Summary of results

The work reported in this dissertation contributes to a contemporary research program in neuroeconomics that identifies factors, which could potentially compromise decision quality.

In our first study, using a well-established social stress induction protocol, we found strong evidence against a temporally dynamic effect of acute stress on revealed preference consistency in a food choice task. Stressed and non-stressed participants showed comparable levels of choice consistency in two time-windows, which is in line with a previous study (Cettolin et al., 2020). The ability to comply with incentive structures under acute stress might also serve as an explanation of divergent observations of stress effects in other choice domains such as social choice (Faber & Häusser, 2021). In addition, further exploration tentatively suggested that revealed preference consistency might be impaired with increasing levels of chronic stress, which we are currently trying to confirm in an ongoing investigation.

In our second study, we used a novel multi-attribute visual choice (MAVC) paradigm, that allowed us to transfer revealed preference theory to visual similarity-based decisions. Specifically, we experimentally tested the influence of memory retrieval of exemplar stimuli, as a model of choice goals, on the choice consistency of similarity-based preference decisions. Memory retrieval was manipulated by varying the retention time between exemplar presentation and choice. Results showed strong evidence against our hypothesis that revealed preference consistency decreases with increasing retention time. However, quality controls indicated that the choice consistency level of our participants was non-discernable from random behavior. Based on oral feedback of participants and further exploration we attributed this to too difficult visual object discrimination in the choice set. Therefore, we deem our results only interpretable to a limited extent regarding our initial hypothesis. However, our paper lays the theoretical foundation and has important implications for the design of further studies on revealed visual preference consistency.

Next, we set out to answer, the question whether economic rationality is malleable by *any* external or internal factors granted everything else equal. Using an unsystematic, qualitative approach to review the literature specifically for revealed preference consistency in a strict sense, we found that for many possible factors there is only ambiguous or no evidence of influence (Study 3). However, after extending the scope for a more liberal definition of consistency, a p-curve analysis yielded clear evidence that at least some factors indeed compromise rationality (Study 4). Still, given the current state of research, it is difficult to draw synthesized conclusions beyond such general statements, which is aggravated by the fact that choice consistency research is conceptually diverse and only loosely connected as indicated by bibliometric indicators. Further, in the interpretation of our results, the findings of our fifth study must be taken into consideration as we aggregated evidence across a wide range of research designs (see below).

In our fifth study, we investigated the reliability of revealed preference consistency in the domain of social decisions. First, in line with previous studies, we found that overall choice consistency was relatively high, and that self-reported decision strategies had intuitive appeal. However, our results also indicated generally poor inter-method reliability and at best moderate test-retest reliability, which is devastating for the identification of replicable relationships with other variables (Hedge et al., 2018). Specifically, endeavors to correlate choice consistency to demographic variables (Choi et al., 2014, American Economic Review), education level (Kim et al., 2018, Science) or brain structures (Chung et al., 2017, The Journal of Neuroscience) might not be fully interpretable, despite wide reception of the aforementioned publications.

From our finding in Study 5, we can derive the hypothesis of low replicability of influence factors on choice consistency, especially for correlational designs. This is, tentatively, in line with our findings from Study 3 (see table 2 of Study 3), which indicated generally mixed results for various influence factors, despite the typically large sample sizes deployed in economic panel studies. However, it appears likely that the true extent of the problem is still shrouded by the lack of replications and publication bias.

Taken together, the work in this dissertation subjected economic rationality to a stress test using multiple approaches. Our findings tentatively suggest that choice consistency is neither a robust nor reliable trait of decision makers, but our empirical work also highlights that not every nuisance (i.e. acute stress) must immediately lead to reduced rationality. Further, our work highlights that economic concepts ought not be naively mistaken for psychometric variables for theoretical (see below) and practical reasons.

A note on the relevance of expected utility theory

In the context of this dissertation, it appears necessary to address the elephant in the room that is the arguable relevance of continued work on EUT.

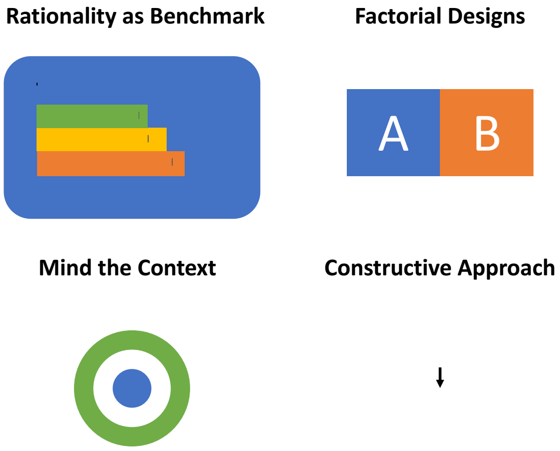
From a descriptive perspective, people's decisions are, generally, not perfectly consistent. Consistency principles must be relaxed significantly in order to capture actual behavior (Rieskamp et al., 2006), and it is still an open discussion how expected utility is represented (neuro-)psychologically, if at all (Hayden & Niv, 2020).

From a predictive perspective, the predictive accuracy of EUT is comparatively lower than for more complex models across multiple studies in different contexts, and especially for larger datasets (Abellan-Perpiñan et al., 2009; de Moraes Ramos et al., 2011; Harless & Camerer, 1994; Peterson et al., 2021), which can be explained by its strict parsimony and disregard of the choice context (Plonsky & Erev, 2021). Concretely, expected utility models show only moderate out-of-sample predictive performance (about 67% correct in binary choice, compared to 65% for simple expected value maximization), even in the context of a low-dimensional experimental task (Garagnani, 2020).

Lastly, from a normative perspective one might argue that we cannot require decision-makers to do what is not (neuro-psychologically) possible (ought implies can; see Foley, 1993, pp. 159–160, for an exemplary argument), and that EUT is too ambiguous to derive prescriptions from in practice (Tversky, 1975).

However, it is also worthwhile to note that EUT, despite its strict parsimony, can account for many regularities of choice behavior, deviations are often of limited severity in terms of costs (Choi et al., 2007), and simple extensions (such reference-dependence and divisive normalization) can explain many systematic deviations. In the domain of predictions, EUT can likely still remain relevant in certain niches that are problematic for more data-intensive technology, e.g. in where the available information is sparse. Generally, "[w]e cannot declare a single winner among theories - much as we cannot declare a best ice cream or university - because the best theory depends on one's tradeoff between parsimony and fit." (Harless & Camerer, 1994, p. 1285). But also, beyond sparse information availability, there are imaginable scenarios where EUT might outperform other theories. It is ultimately an empirical question to identify conditions under which EUT performs well, predictively. Lastly, EUT still has wide traction as a normative theory in economics and decision analysis (Małecka, 2020) and despite various criticisms, no outstanding successor theory has established itself so far (Moscati, 2016). Further, it should be noted that criticisms often do not question the normative plausibility of the choice axioms (e.g. the independence axiom), and that expected utility is still considered a "rational bedrock" (Briggs, 2019).

Figure 3: Guidelines for Future Research on Economic Rationality



Based on theoretical and empirical considerations in this dissertation, four guidelines for future research on economic rationality can be derived that that in the thought tradition of scientific instrumentalism. First, EUT rationality should be treated as a normative benchmark, not as a descriptive model of decision-making. As in our empirical work, EUT can be used as a normative benchmark, for example to identify conditions under which decision-makers' (relative) rationality deteriorates. Second, due to reliability issues in contemporary measurements of rationality that we have identified in Study 5, research should utilize factorial rather than correlational designs. Third, when drawing conclusions based on empirical data we must consider the context of measurement and evaluate the validity of our ad-hoc assumptions (e.g. via domain expertise or self-reports). Fourth and lastly, theoretical work should advance past pure demonstrations of the fallibility of EUT and engage in a constructive theoretical approach.

In any case, a general note of caution seems in place to be given to the current generation of researchers working on EUT (such as myself). At this point, it is absolutely necessary to specify which interpretation of EUT – descriptive, predictive, or normative – is being used to derive experimental hypotheses from, for else the interpretation of any results is ambiguous as well. For example, as outlined above, it is common grounds across disciplines and for multiple years, that descriptive EUT, in its typical application, is inaccurate; pure demonstrations of its descriptive shortcomings without any constructive contribution appear unnecessary (as I once learned from an anonymous editor). This criticism, however, does not apply if we are intrinsically interested in the influence of some factor on decision quality in comparison to the benchmark of economic rationality, e.g. to identify risk factors in the decision-making of practitioners, like it is the case for the work reported in this dissertation. I believe that this framing of EUT as a *tool*, i.e. the instrumentalist perspective in the philosophy of science, is also most suitable for the interpretation of our empirical work (see Figure 3).

Methodological limitations

Beyond placing our work in the long-standing debate around EUT, I must also address some methodological concerns.

Generally, in our experimental work we made the typical *ceteris paribus* assumption: we assumed that inconsistencies did not arise due to uncontrolled changes in variables that are relevant to the decision process. For Study 1, we facilitated this condition by placing our behavioral tests in distinct neurohormonal time windows of the stress response, as deducted from theoretical work and checked via hormonal assessment. In addition, in Studies 1, 2 and 5 we took care to prevent choice repetition effects, by either fully or partially randomizing choice options. Lastly, in Studies 1, 2 and 5 our consistency tests generally did not last more than a few minutes each. For Studies 3 and 4 we rely on the diligence of the original experiments in the literature. However, theoretically, it is still almost always possible to construe a post-hoc rationalization of inconsistencies due to changes in decision-relevant variables, which are, strictly speaking, unknown to us (Edwards, 1954; Regenwetter & Davis-Stober, 2012 give a very illustrative example on the overt inconsistency of a PhD advisor). Such data aggregation artifacts (Regenwetter et al., 2011) pose a challenge to the validity of choice consistency tests as measurements of rationality. Fortunately, the severity of this challenge to the interpretability of our experimental work is limited. As in all of our studies we are making only relative and comparative statements of rationality, aggregation artifacts can be neglected as long as they are not linked to experimental conditions or predictor variables. A priori, the existence of such a link appears the most likely for Study 1 due to dynamic changes of preferences in the stress group. However, the presence of this (as of any other) group effect is not supported by the data, given strong evidence for a null result. As for our literature work, however, a final evaluation of this problem is naturally difficult – and goes to show the importance of creating awareness of aggregation artifacts in experimentalists. A second concern of internal validity arises from the results of our own Study 5: given that choice consistency measures show low reliability, correlational designs such as in Study 2 and many original studies included in Studies 3 and 4 are problematic (Hedge et al., 2018) – future research must show the consequences of this (late) realization.

A concern regarding the generalizability of our results arises given the limited representativeness of our participant samples, mostly recruited from the local student body and a click worker platform. This is not only a platitude, given that we (see Study 3) and others (Choi et al., 2014) found evidence that demographic variables could affect choice. Currently, it is unclear whether there are also interaction effects with other variables, as demographic variables are often neglected or only considered as linear covariates in statistical models. Hence, we can neither confirm nor rule out whether our results would hold up in a more representative sample. A second concern of generalizability arises given that we only conducted laboratory (or online) experiments. Especially if the goal of research on EUT is to be relevant the practitioner disciplines it is necessary to move away from 'lab-only' research (as already argued by Kahneman, 1991).

Our literature work, specifically, might also be affected by typical biases in the scientific literature. While our P-Curve analysis (Study 4) tentatively suggests that findings in the literature are not driven by publication bias or p-hacking, a critical reader of recent publications notices that transparency, open data and publications of null results are still lacking in the field, which is aggravated by the ongoing hunt for 'spectacular' deviations from rationality. Therefore, in our experimental work, we tried our best to follow open scientific practice (see Open Science statement).

Lastly, it must be made clear that all of the work reported in this dissertation trivially does not allow for final conclusions – despite demarking the finale of my doctoral studies – but should rather be framed as a contribution to the scientific evidence accumulation process.

Outlook for future research

An exploratory result in Study 1 tentatively suggested a negative relationship of chronic stress and choice consistency. However, due to the nature of our design we could only perform a correlational analysis of choice consistency with a self-report measure of chronic. Future research should provide a confirmatory account of this hypothesis, ideally using more valid measures of chronic stress (e.g. biological markers). It could also be interesting to translate the paradigm to an animal model similar to Hu et al. (2021), which would allow an experimental intervention of chronic or early-life stress and a better understanding of the neurobiological underpinnings (Cameron & Schoenfeld, 2018; Friedman et al., 2017).

Motivated by new methodological insights from Study 2, I believe that experiment should be repeated, drawing from the acquired learnings with an improved design. Specifically, a relatively consistent baseline choice behavior must be established via intensive piloting and perhaps individualized (and adaptive) difficulty adjustment of the task. Further, it might be useful to rely on a more easy to process and validated class of stimuli (Lebaz et al., 2020; Liu &

Kersten, 1998). Lastly, work on the connection of economic and psychophysical concepts akin to Dziewulski (2020) appears to be an interesting interdisciplinary theoretical perspective.

More generally, since decision analysis and applied economics seem to make up the largest part of the actual userbase of EUT, further research and development could start with the identification of problems, challenges and questions that arise in practice. Concretely, the theoretically well-founded integration of domain expertise within EUT could begin with a better understanding of how practitioners perceive and cognize typical problems. Here, EUT research could look for inspiration in the field of naturalistic decision-making (Hoffman & Klein, 2017; Klein, 2015; Walker, 2017) and cognitive task analysis, specifically, which specifically targets the needs of practitioners in theory development. Further, if the field manages to move from small-sample lab experiments to larger scale behavior measurements, neural network models might be used to uncover multi-dimensional representations of choice objects as well as improve predictive accuracy (for some interesting applications see: Hebart et al., 2020; Ma & Peters, 2020; Peterson et al., 2021).

Lastly, reflecting on the theoretical criticism of EUT outlined above, interestingly, similar arguments emerge from multiple perspectives. For example, we can neither describe, predict or normatively evaluate decisions regarding their subjective rationality if we completely disregard the subjective representation of choice objects, as well as the environment and context of the decision-maker. While this is not a challenge to the logical consistency of EUT, it shows that the theory needs to be extended to be actually applicable to data. This short-coming is indirectly reflected both, in the neuroeconomic pursuit of complementing EUT with insights from neuroscience and psychology, as well as decision analysis infusing domain knowledge, both in an attempt to bridge the gap between theory and data.

Another example, is the limited neuroeconomic realism of EUT (e.g. van Rooij et al., 2018) which is, of course, a direct issue from a descriptive perspective, but also curtails the normative force of the theory if we follow the *ought implies can* principle. We must ask ourselves, intuitively, is it rational to chase a utopia (Foley, 1993; Zynda, 1996)?

Conclusion

This dissertation reports on a series of studies stress-testing the concept of economic rationality from multiple perspectives. Our findings contribute to the emerging picture that relative choice consistency is neither a robust nor reliable trait of decision makers. Beyond the reported studies, this dissertation provides a theoretical embedding – specifically the instrumental perspective (see figure 3) – which is indispensable to the utility of contemporary research on EUT. Future research is still and always indebted to address challenges of external validity and generalizability (Kahneman, 1991), to more transparency, as well as to the development of a theoretical successor of EUT.

Open Science Statement

All data and code to reproduce the results reported in this dissertation have been made publicly available on the Open Science Framework. In addition, for Study 1, we created a Data in Brief publication (Study 1b) to facilitate the reuse of our resources. With the exception of Study 1, where journal guidelines prohibited this, all manuscripts have been published as preprints. Study 2 has been performed as a registered report. Finally, 2 out of 5 studies reported in this dissertation describe null effects.

References

Abellan-Perpiñan, J. M., Bleichrodt, H., & Pinto-Prades, J. L. (2009). The predictive validity of prospect theory versus expected utility in health utility measurement. *Journal of Health Economics*, *28*(6), 1039–1047. https://doi.org/10/djgbbm

Afriat, S. N. (1972). Efficiency Estimation of Production Functions. *International Economic Review*, *13*(3), 568–598. JSTOR. https://doi.org/10/bkn49z

Afriat, S. N. (1973). On a system of inequalities in demand analysis: An extension of the classical method. *International Economic Review*, 460–472. https://doi.org/10/fdr8kn

Arnsten, A. F. (2009). Stress signalling pathways that impair prefrontal cortex structure and function. *Nature Reviews Neuroscience*, *10*(6), 410. https://doi.org/10/dn9bvx

Arnsten, A. F., Wang, M. J., & Paspalas, C. D. (2012). Neuromodulation of thought: Flexibilities and vulnerabilities in prefrontal cortical network synapses. *Neuron*, *76*(1), 223–239. https://doi.org/10/f4bzzf

Beilock, S. L., & DeCaro, M. S. (2007). From poor performance to success under stress: Working memory, strategy selection, and mathematical problem solving under pressure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(6), 983–998. https://doi.org/10/fqdbfc

Bentham, J. (1780). *An introduction to the principles of morals and legislation*. Dover Publications.

Brand, M., Fujiwara, E., Borsutzky, S., Kalbe, E., Kessler, J., & Markowitsch, H. J. (2005). Decision-Making Deficits of Korsakoff Patients in a New Gambling Task With Explicit Rules: Associations With Executive Functions. *Neuropsychology*, *19*(3), 267–277. https://doi.org/10/cqqkf3

Brand, M., Schiebener, J., Pertl, M.-T., & Delazer, M. (2014). Know the risk, take the win: How executive functions and probability processing influence advantageous decision making under risk conditions. *Journal of Clinical and Experimental Neuropsychology*, *36*(9), 914–929. Scopus. https://doi.org/10/gk9ptv

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

Cameron, H. A., & Schoenfeld, T. J. (2018). Behavioral and structural adaptations to stress. *Frontiers in Neuroendocrinology*, *49*, 106–113. https://doi.org/10/gdp4z6

Cettolin, E., Dalton, P. S., Kop, W. J., & Zhang, W. (2020). Cortisol meets GARP: The effect of stress on economic rationality. *Experimental Economics*, *23*(2), 554–574. https://doi.org/10.1007/s10683-019-09624-z

Choi, S., Fisman, R., Gale, D., & Kariv, S. (2007). Consistency and heterogeneity of individual behavior under uncertainty. *American Economic Review*, *97*(5), 1921–1938. https://doi.org/10/c3665n

Choi, S., Kariv, S., Müller, W., & Silverman, D. (2014). Who Is (More) Rational? *American Economic Review*, *104*(6), 1518–1550. https://doi.org/10/76w

Chung, H.-K., Tymula, A., & Glimcher, P. (2017). The Reduction of Ventrolateral Prefrontal Cortex Grey Matter Volume Correlates with Loss of Economic Rationality in Aging. *The Journal of Neuroscience*, *37*(49), 1171–17. https://doi.org/10/gcq3j8

Corner, J. L., & Kirkwood, C. W. (1991). Decision Analysis Applications in the Operations Research Literature, 1970–1989. *Operations Research*, *39*(2), 206–219. https://doi.org/10/bwk6qm

Cubitt, R. P., & Sugden, R. (2001). On money pumps. *Games and Economic Behavior*, *37*(1), 121–160. https://doi.org/10/cxjbgc

De Kloet, E. R., Joëls, M., & Holsboer, F. (2005). Stress and the brain: From adaptation to disease. *Nature Reviews Neuroscience*, *6*(6), 463. https://doi.org/10/b8r9nv

de Moraes Ramos, G., Daamen, W., & Hoogendoorn, S. (2011). Expected Utility Theory, Prospect Theory, and Regret Theory Compared for Prediction of Route Choice Behavior. *Transportation Research Record: Journal of the Transportation Research Board*, *2230*(1), 19–28. https://doi.org/10/ckmxtq

Droste, S. K., de Groote, L., Atkinson, H. C., Lightman, S. L., Reul, J. M., & Linthorst, A. C. (2008). Corticosterone levels in the brain show a distinct ultradian rhythm but a delayed response to forced swim stress. *Endocrinology*, *149*(7), 3244–3253.

Dziewulski, P. (2020). Just-noticeable difference as a behavioural foundation of the critical cost-efficiency index. *Journal of Economic Theory*, *188*, 105071. https://doi.org/10/gmv6gm

Edwards, W. (1954). THE THEORY OF DECISION MAKING. *Psychological Bulletin*, *51*(4), 38. https://doi.org/10/crnw6p

Evanson, N. K., Tasker, J. G., Hill, M. N., Hillard, C. J., & Herman, J. P. (2010). Fast feedback inhibition of the HPA axis by glucocorticoids is mediated by endocannabinoid signaling. *Endocrinology*, *151*(10), 4811–4819.

Faber, N. S., & Häusser, J. A. (2021). Why stress and hunger both increase and decrease prosocial behaviour. *Current Opinion in Psychology*. https://doi.org/10/gmnqxs

Fink, G. (2016). Chapter 1 - Stress, Definitions, Mechanisms, and Effects Outlined: Lessons from Anxiety. In G. Fink (Ed.), *Stress: Concepts, Cognition, Emotion, and Behavior* (pp. 3–11). Academic Press. https://doi.org/10.1016/B978-0-12-800951-2.00001-7

Foley, R. (1993). *Working without a net: A study of egocentric epistemology*. Oxford University Press on Demand.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. https://doi.org/10/b98rhb

Friedman, A., Homma, D., Bloem, B., Gibb, L. G., Amemori, K., Hu, D., Delcasso, S., Truong, T. F., Yang, J., Hood, A. S., Mikofalvy, K. A., Beck, D. W., Nguyen, N., Nelson, E. D., Toro Arana, S. E., Vorder Bruegge, R. H., Goosens, K. A., & Graybiel, A. M. (2017). Chronic Stress Alters Striosome-Circuit Dynamics, Leading to Aberrant Decision-Making. *Cell*, *171*(5), 1191-1205.e28. https://doi.org/10/gcmnm5

Gabaix, X., & Laibson, D. (2017). *Myopia and discounting*. National bureau of economic research. https://www.nber.org/papers/w23254

Garagnani, M. (2020). *The Predictive Power of Risk Elicitation Tasks* [Working Paper]. Department of Economics Working Paper Series. https://www.zora.uzh.ch/id/eprint/190318/

Gathmann, B., Schulte, F. P., Maderwald, S., Pawlikowski, M., Starcke, K., Schäfer, L. C., Schöler, T., Wolf, O. T., & Brand, M. (2014). Stress and decision making: Neural correlates of the interaction between stress, executive functions, and decision making under risk. *Experimental Brain Research*, *232*(3), 957–973. https://doi.org/10/f5s367

Groeneweg, F. L., Karst, H., de Kloet, E. R., & Joëls, M. (2011). Rapid non-genomic effects of corticosteroids and their role in the central stress response. *Journal of Endocrinology*, *209*(2), 153–167. https://doi.org/10/bckq3j

Gul, F., & Pesendorfer, W. (2008). The case for mindless economics. *The Foundations of Positive and Normative Economics: A Handbook*, *1*, 3–42.

Harless, D. W., & Camerer, C. F. (1994). The Predictive Utility of Generalized Expected Utility Theories. *Econometrica*, *62*(6), 1251. https://doi.org/10/b462rb

Hayden, B., & Niv, Y. (2020). *The case against economic values in the brain* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/7hgup

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185. https://doi.org/10/ghfkgv

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10/gddfm4

Hermans, E. J., Henckens, M. J., Joëls, M., & Fernández, G. (2014). Dynamic adaptation of large-scale brain networks in response to acute stressors. *Trends in Neurosciences*, *37*(6), 304–314. https://doi.org/10/gg4tjf

Hermans, E. J., van Marle, H. J., Ossewaarde, L., Henckens, M. J., Qin, S., van Kesteren, M. T., Schoots, V. C., Cousijn, H., Rijpkema, M., & Oostenveld, R. (2011). Stress-related noradrenergic activity prompts large-scale neural network reconfiguration. *Science*, *334*(6059), 1151–1153. https://doi.org/10/dbf5vb

Hinz, B., & Hirschelmann, R. (2000). Rapid Non-Genomic Feedback Effects of Glucocorticoids on CRF-Induced ACTH Secretion in Rats. *Pharmaceutical Research*, *17*(10), 1273–1277. https://doi.org/10/dvb9m2

Hoffman, R. R., & Klein, G. L. (2017). Challenges and Prospects for the Paradigm of Naturalistic Decision Making. *Journal of Cognitive Engineering and Decision Making*, *11*(1), 97–104. https://doi.org/10/ggskzq

Horkheimer, M. (1946). *Eclipse of reason*. Institute of Social Research, Columbia University.

Houthakker, H. S. (1950). Revealed preference and the utility function. *Economica*, *17*(66), 159–174. https://doi.org/10/fhq8cj

Hu, Y., Nitsch, F. J., van Wingerden, M., & Kalenscher, T. (2021). *Cross-species comparison of human and rodent primary reward consumption under budget constraints* [Preprint]. bioRxiv. https://www.biorxiv.org/content/early/2021/07/29/2021.07.28.454138

Huang, I. B., Keisler, J., & Linkov, I. (2011). Multi-criteria decision analysis in environmental sciences: Ten years of applications and trends. *Science of The Total Environment*, *409*(19), 3578–3594. https://doi.org/10/cm6qzj

Hume, D. (1738). *A treatise of human nature: Being an attempt to introduce the experimental method of reasoning into moral subjects*. Collins.

Joëls, M., & Baram, T. Z. (2009). The neuro-symphony of stress. *Nature Reviews Neuroscience*, *10*(6), 459. https://doi.org/10/fbhvrr

Johnson, E. J., Häubl, G., & Keinan, A. (2007). Aspects of endowment: A query theory of value construction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(3), 461–474. https://doi.org/10/dcgsrs

Kahneman, D. (1991). Article Commentary: Judgment and Decision Making: A Personal View. *Psychological Science*, *2*(3), 142–145. https://doi.org/10/cg53mb

Keefer, D. L., Kirkwood, C. W., & Corner, J. L. (2004). Perspective on Decision Analysis Applications, 1990–200. *Decision Analysis*, *1*(1), 34.

Kim, H. B., Choi, S., Kim, B., & Pop-Eleches, C. (2018). The role of education interventions in improving economic rationality. *Science*, *362*(6410), 83–86. https://doi.org/10.1126/science.aar6987

Klein, G. (2015). A naturalistic decision making perspective on studying intuitive decision making. *Journal of Applied Research in Memory and Cognition*, *4*(3), 164–168. https://doi.org/10/gfgssb

Lebaz, S., Sorin, A.-L., Rovira, K., & Picard, D. (2020). Widgets: A new set of parametrically defined 3D objects for use in haptic and visual categorization tasks. *European Review of Applied Psychology*, *70*(3), 100552. https://doi.org/10/gms4z8

Liu, Z., & Kersten, D. (1998). 2D observers for human 3D object recognition? *Vision Research*, *38*(15–16), 2507–2519. https://doi.org/10/b3jx3b

Ma, W. J., & Peters, B. (2020). *A neural network walks into a lab: Towards using deep nets as models for human behavior* [Preprint]. arXiv. http://arxiv.org/abs/2005.02181

Maier, S. U., Makwana, A. B., & Hare, T. A. (2015). Acute stress impairs self-control in goal-directed choice by altering multiple functional connections within the brain's decision circuits. *Neuron*, *87*(3), 621–631. https://doi.org/10/f7qjmj

Małecka, M. (2020). The normative decision theory in economics: A philosophy of science perspective. The case of the expected utility theory. *Journal of Economic Methodology*, *27*(1), 36–50. https://doi.org/10/gmk8t3

Margittai, Z., Nave, G., Strombach, T., van Wingerden, M., Schwabe, L., & Kalenscher, T. (2016). Exogenous cortisol causes a shift from deliberative to intuitive thinking. *Psychoneuroendocrinology*, *64*, 131–135. https://doi.org/10/gfzkbp

Mill, J. S. (1863). *Utilitarianism*. Cleveland: Cambridge University Press.

Morgenstern, O., & Von Neumann, J. (1953). *Theory of games and economic behavior*. Princeton university press.

Moscati, I. (2016). Retrospectives: How Economists Came to Accept Expected Utility Theory: The Case of Samuelson and Savage. *Journal of Economic Perspectives*, *30*(2), 219–236. https://doi.org/10/ghbg8w

Nitsch, F. J., & Kalenscher, T. (2020). *Keeping a cool head at all times. What determines choice consistency?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/etyhx

Nitsch, F. J., & Kalenscher, T. (2021a). *Influence of memory processes on choice consistency.* [Preprint]. PsyArXiv. psyarxiv.com/74br5

Nitsch, F. J., & Kalenscher, T. (2021b). *How robust is rational choice?* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/zv2m8

Nitsch, F. J., Lüpken, L. M., Lüschow, N., & Kalenscher, T. (2021). *Inconsistently consistent: Rationality is not reliable* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/gd9zs

Nitsch, F. J., Sellitto, M., & Kalenscher, T. (2021a). Trier social stress test and food-choice: Behavioral, self-report & hormonal data. *Data in Brief*, *37*, 107245. https://doi.org/10/gmd57p

Nitsch, F. J., Sellitto, M., & Kalenscher, T. (2021b). The effects of acute and chronic stress on choice consistency. *Psychoneuroendocrinology*, *131*, 105289. https://doi.org/10/gk9pvk

Peterson, J. C., Bourgin, D. D., Agrawal, M., Reichman, D., & Griffiths, T. L. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214. https://doi.org/10/gkh2nn

Plonsky, O., & Erev, I. (2021). To predict human choice, consider the context. *Trends in Cognitive Sciences*. https://doi.org/10/gmkdtd

Qin, S., Hermans, E. J., van Marle, H. J. F., Luo, J., & Fernández, G. (2009). Acute Psychological Stress Reduces Working Memory-Related Activity in the Dorsolateral Prefrontal Cortex. *Biological Psychiatry*, *66*(1), 25–32. https://doi.org/10/cc4v2s

Ramsey, F. (1926). Truth and probability. In A. Eagle (Ed.), *Philosophy of probability: Contemporary readings* (pp. 52–94). Routledge.

Regenwetter, M., Dana, J., & Davis-Stober, C. P. (2011). Transitivity of preferences. *Psychological Review*, *118*(1), 42. https://doi.org/10/b3fd9v

Regenwetter, M., & Davis-Stober, C. P. (2012). Behavioral Variability of Choices Versus Structural Inconsistency of Preferences. *Psychological Review*, *119*(2), 408–416. https://doi.org/10/f4rvgr

Rieskamp, J., Busemeyer, J. R., & Mellers, B. A. (2006). Extending the Bounds of Rationality: Evidence and Theories of Preferential Choice. *Journal of Economic Literature*, *44*(3), 631–661. https://doi.org/10/dzqttm

Samuelson, P. A. (1938). A note on the pure theory of consumer's behaviour. *Economica*, *5*(17), 61–71. https://doi.org/10/cqz9wq

Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10/gffnn9

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). *p* -Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. *Perspectives on Psychological Science*, *9*(6), 666–681. https://doi.org/10/f6q2j6

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*(6), 1146–1152. https://doi.org/10/gf5smq

Tasker, J. G., Di, S., & Malcher-Lopes, R. (2006). Rapid glucocorticoid signaling via membrane-associated receptors. *Endocrinology*, *147*(12), 5549–5556.

Tversky, A. (1975). A critique of expected utility theory: Descriptive and normative considerations. *Erkenntnis*, *9*(2). https://doi.org/10/c7ztjd

Ulrich-Lai, Y. M., & Herman, J. P. (2009). Neural regulation of endocrine and autonomic stress responses. *Nature Reviews Neuroscience*, *10*(6), 397. https://doi.org/10/cgk3tz

van Rooij, I., Wright, C. D., Kwisthout, J., & Wareham, T. (2018). Rational analysis, intractability, and the prospects of 'as if'-explanations. *Synthese*, *195*(2), 491–510. https://doi.org/10/gg84h6

Varian, H. R. (1982). The nonparametric approach to demand analysis. *Econometrica: Journal of the Econometric Society*, 945–973. https://doi.org/10/b6s3sx

Varian, H. R. (2006). Revealed preference. *Samuelsonian Economics and the Twenty-First Century*, 99–115. https://doi.org/10/cbssvr

Velasquez, M., & Hester, P. T. (2013). *An Analysis of Multi-Criteria Decision Making Methods*. *10*(2), 11.

Vineberg, S. (2016). Dutch book arguments. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Spring 2016). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2016/entries/dutch-book/

Walker, R. (2017). Naturalistic research. *Research Methods & Methodologies in Education*, 78–84.

Wimmer, G. E., & Shohamy, D. (2012). Preference by Association: How Memory Mechanisms in the Hippocampus Bias Decisions. *Science*, *338*(6104), 270–273. https://doi.org/10/f39gxj

Zynda, L. (1996). Coherence as an ideal of rationality. *Synthese*, *109*(2), 175–216. https://doi.org/10/bfrdnr

Appendix A – The Dutch book argument

The Dutch book argument (Ramsey, 1926) has the following structure:

Preposition 1. Rationality excludes non-sensical behavior.

Preposition 2. If subjective probabilities do not follow the axioms of probability, this allows non-sensical behavior.

Conclusion: Hence, rationality requires subjective probabilities to follow the axioms of probability (i.e. normality, non-negativity, additivity for mutually exclusive events).

I shall provide a partial proof of Preposition 2 for the axiom of additivity, that is $p(H) + p(T) = p(H \vee T)$.

Let us assume a gambler names honest prices for which they are accepting to both, take and offer bets on a coin flip against the house. We can imagine that the honest price corresponds the product of the utility of the potential earning $U(x)$ and the corresponding subjective probability p of the outcome (or a homogeneous function thereof). Let us assume that there are three bets: a bet on H, that wins x if the coin shows heads, a bet on T, that wins x if the coin shows tails, and a bet on $H \vee T$, that wins x if the coin shows either heads or tails, for which the gambler assigns the prices $h = p(H) \times U(x)$, $t = p(T) \times U(x)$, and $c = p(H \vee T) \times U(x)$, respectively. If the gambler would then agree to offer bets on H and T and take a bet on $H \vee T$, each at their fair price, then, for any outcome of the coin flip, they would earn $h + t - c$ and if $c > h + t$, they would lose to the house (Dutch book). If the gambler would conversely agree to take bets on H and T and offer a bet on $H \vee T$, each at their fair price. Then, for any outcome, they earn $c - r - t$ and if $c < h + t$, they would lose to the house as well. Hence, $c = h + t$ and, thus, $p(H) + p(T) = p(H \vee T)$ must hold, else it allows non-sensical behavior. ■