# COMPUTATIONAL AND BEHAVIOURAL PRINCIPLES UNDERLYING REACTIONS TO SOCIAL REWARDS

Filip Gesiarz

Prepared under the supervision of:

Professor Tali Sharot

Submitted for the degree of Doctor of Philosophy

Department of Experimental Psychology

University College London

January 2020

## Declaration

I, Filip Gesiarz, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

## Abstract

Social contexts often change how people engage with and evaluate available rewards, leading to behaviours that defy simple rules of reward maximization. The current thesis aims to characterize some of the principles that underlie reactions to rewards obtained in a social context and formalize them in computational models.

In study 1, I explore how social reward distributions change the hedonic and motivational value of rewards. The study shows that people are often demotivated and distressed by the unfairness of the distribution, and are less willing to work for their offered rewards even if they are the ones benefiting from the unfair situation. I introduce a model that characterizes the responses to reward distributions as a linear combination of statistical dispersion and rank ordering of the rewards and show that its predictions fit more closely to observed behaviour than many other alternative models suggested in behavioural economics and psychology. In study 2, I test how people form subjective judgments about reward distributions. The study demonstrates that subjective judgments are biased by personal position in the distribution, and violate several normative axioms used in economics. In study 3, I demonstrate the effect of the international distribution of rewards on life-evaluations: the study shows that life evaluations are not only sensitive to comparisons with citizens in one's own country, but also to comparisons with people in other countries. The model characterizing the response to reward distributions as a linear combination of statistical dispersion and rank ordering again is shown to fit well-being data better than any other alternative. Study 4 focuses on the influence of the distribution of beliefs about oneself on preferences for feedback. It shows that people sometimes might prefer negative feedback, and describes heuristics and learning mechanisms that lead to this behaviour.

The four studies presented in this thesis expand our knowledge of how external and internal social contexts change our experience with rewards. They

introduce computational models that aim to formalize such contextual influences, contributing to a more mechanistic understanding of these effects.

## Impact statement

This thesis offers several practical and theoretical insights relevant to our society.

In Chapter 2 I demonstrate that those who are disadvantaged are demotivated by both their position in the society and unfairness of the situation. This effect might be responsible for motivational poverty-traps, in which those who are disadvantaged are also the ones who are the most demotivated to improve their situation. I this experiment, I also test predictions of many competing theories, advancing our theoretical understanding of the influence of context on evaluation of rewards.

In Chapter 3, we show that people's subjective perceptions of inequality deviate from the measures used in economics and propose a measure that more closely approximates such subjective perceptions. We also identify which standard measures most closely correspond to lay perceptions. These findings provide practical solutions for governments aiming to minimize inequality as seen by the public, rather than as operationalized by economists.

In Chapter 4, we identify the most plausible model of contextual influence of international living standards in a representative sample of more than 2 million individuals from 156 countries. We discover that well-being is not only affected by comparisons with people in one's own country but also with people in other countries, as well as inequality between countries. This finding can inform our understanding of factors involved in migration decisions, as well as provides an explanation for tensions within countries that can arise due to factors outside these countries. These insights might be helpful for policy-makers who tackle the issues of immigration and national well-being.

In Chapter 5, we describe computational mechanisms underlying the phenomenon of seeking confirmatory information about oneself. By doing so, we create a bridge between theories in social psychology and a reinforcement learning framework developed in decision neuroscience. Formal

characterization of processes involved in confirmatory information seeking is of relevance for computational psychiatry and can help in designing interventions for anxiety and major depressive disorder in which this mechanism self-reinforces negative views of oneself.

Overall, this thesis provides a more complete understanding of reactions to social rewards, informing political science, economics, psychology, and neuroscience, and proposing a computational framework that can be used as a starting point in future studies.

To my mother, whose personal sacrifices and hard work allowed me to pursue educational opportunities that she never had.

## Acknowledgments

The four years of this Ph.D. have been an incredible intellectual adventure. As any adventure, it had its ups and downs and would have not been the same without the people who I was fortunate to meet and work with during that time.

First of all, I would like to thank Tali for her supervision and for creating many wonderful opportunities throughout the years. The Affective Brain Lab was a truly stimulating place to do a Ph.D. in, and I do believe it made me a much more mature and well-rounded researcher (with a lot still to learn). I would like to thank Molly Crockett for her support and encouragement that gave me much needed confidence before the start of my Ph.D., Peter Dayan for invaluable insights and discussions during my first year, and Jan De Neve for a very fruitful collaboration.

I am very grateful to Neil and Caroline for their kindness, welcoming me in the lab and helping me settle in during my first year; Donal, Eleanor, Seb and Steph for their company during the first half of my Ph.D., as well as their knowledge and helpful comments that influenced many parts of my work. Especially warm thanks to Joe Marks and Chris Kelly, for sharing many laughs, gossips and intellectual battles, and for being not only supportive colleagues but also dear friends. A special mention to Yulin, for her contagious positive energy, as well Laura, Valentina, and Irene, for many positive interactions during my last few months.

A lot of appreciation goes to UCL staff members for creating one of the most smoothly operating universities that I had experience with. Special thanks to John Draper and Jeremy Skipper, who provided invaluable support during the more difficult parts of my Ph.D.

Lastly, I would like to thank my partner in life who has been an endless source of love and joy, fuelling me throughout my Ph.D.

**TABLE OF CONTENTS**

# Chapter 1

# General Introduction

## OVERVIEW

The current thesis explores how different social contexts can change how people react to and engage with rewards. On the one hand, I focus on the external contexts, such as the distribution of wealth in a person's group or society, and investigate how they might influence what a person thinks and feels about their available rewards. On the other hand, I look at the internal contexts, such as distribution of beliefs about oneself, and investigate how they can influence how people respond to and seek rewards in the form of social feedback. In all cases, I put a special emphasis on the formal description of the mechanisms and computational principles through which social contexts influence the evaluation of rewards.

The idea that context influences perception is as old as experimental psychology itself and goes back to the early days of gestalt school of thought (Ehrenfels, 1890). Contextual effects on vision and other sensory domains have been well defined since (Eagleman, 2001; Jäkel et al., 2016). In contrast, we know little about how social context changes the subjective experience with rewards, with many questions remaining unanswered: What are the computational principles that determine how social context changes the value of rewards? How do the properties of reward distributions in a population influence the hedonic experience with them? How do beliefs about oneself and internalized norms reshape the reward evaluation functions? These and other related questions are addressed in four studies.

In chapter two, I investigate how the motivational and hedonic value of an offered reward is changed by the context of rewards offered to others in the group. The study described in this chapter decomposes two aspects of inequality of rewards that have often been interpreted as a single phenomenon: relativity of rewards and unfairness of the distribution. It tests how these two factors influence momentary well-being and motivation to pursue rewards in an experiment in which participants have to repeatedly decide if they want to work for randomly drawn rewards while observing what

rewards were offered to others. I hypothesized that those with relatively lower rewards, and groups with more unfair distributions of rewards, will be less motivated to work for their rewards, even if the absolute value of their offered reward is the same as in the case of relatively higher rewards and more fair distributions respectively.

In chapter three, I take a closer look at the mechanisms of perception of inequality. The study described in this chapter investigates how people form explicit judgments about the distributions of rewards. In an experiment, I present people with a selection of 60 different reward distributions and asked them about how equal they are. The study verifies to what extent lay perceptions of inequality agree with normative axioms assumed in economics. I hypothesized that lay perceptions of inequality will violate several axioms used in economics in the quantification of statistical dispersion, and that self-interested biases will influence objective judgments of inequality

In chapter four, I return to the question of how the hedonic value of offered rewards is changed by the context of rewards offered to others in the society, and investigate to what extent findings from chapter two are echoed in the international well-being reports. The study described in this chapter reports an analysis of responses from a large-scale survey containing data points from over two million individuals. It examines how inequality between countries and the relative position of a country can change how people evaluate their living standards. Additionally, it discerns the influences of the group and personal identities on such comparisons. I hypothesized that citizens in countries with a lower relative international position, and regions with more inequality between countries, will have lower life satisfaction, despite having a similar living standard as citizens in countries with a relatively higher position on the international stage and regions with less inequality between countries.

In chapter five, I focus on the effect of the internal context of one's beliefs about oneself on motivational and hedonic value of social feedback. In

an experiment in which participants are repeatedly provided with an opportunity to reveal higher or lower evaluations of them made by others, I investigate what drives people to look for confirmatory information about oneself, and how beliefs about oneself can change how people respond to positive and negative information. The study tests to what extent the observed behaviour could be characterized by several different decision heuristics and learning algorithms. I hypothesized that confirmatory choices would be followed by an increased mood and certainty about one's self-evaluation in comparison to disconfimatory choices; and that different people would employ either heuristics or learning mechanisms depending on the discrepancy of their self-views and provided feedback.

The next few sections provide an overview of concepts and theories relevant to studying the contextual effects involved in responses to social rewards.

## INEQUALITY OF REWARDS AND SOCIAL PREFERENCES

Do people prefer equal distribution of rewards in society? The existence of pure other-regarding preferences seemingly defies the axiom of rational self-interest assumed in neoclassical economics. Nevertheless, acting with regard for the well-being of others is widespread across cultures and found in the animal kingdom (Waal, 1997; Henrich et al., 2001; Engel, 2011). One of the most well-known descriptive models of these behaviours in economics is the inequality aversion model, according to which both advantageous inequality (being better off than others) and disadvantageous inequality (being worse off than others) have a negative utility (Fehr and Schmidt, 1999). In this model, advantageous and disadvantageous inequalities are conceptualized as a weighted sum of differences between all incomes lower or higher than one's own income, respectively.

Inequality-aversion model is mainly supported by studies using dictator games, in which people granted initial endowment often decide to re-

distribute some part of it to participants who did not receive any endowment (Engel, 2011; Bechtel et al., 2018). Consistently with this model, a study employing computational modelling has demonstrated that advantageous and disadvantageous inequality have indpendent negative effects on momentary subjective well-being – suggesting that these two types of inequality not only influence prosocial behaviours but also change the experience with rewards (Rutledge et al., 2016). Differential experience of unequal rewards is also supported by an fMRI study that found that rewards contributing to a decrease of inequality are associated with a stronger activation of value-related brain regions than rewards contributing to an increase of inequality (Tricomi et al., 2010).

The inequality-aversion hypothesis has been criticized on many grounds. On the one hand, some experiments have shown that framing the experiment as a game that allows people to take money from others can cause inequality-seeking, suggesting that re-distribution behaviours might be more regulated by social expectations rather than actual preferences (List, 2007; Bardsley, 2008; but see: Bechtel et al., 2018). This interpretation of inequality-aversion has been also suggested before in a guilt-aversion model, according to which people try to minimize feeling guilty after failing to fulfil social expectations (Battigalli and Dufwenberg, 2007; Nihonsugi, Ihara and Haruno, 2015; van Baar, Chang, and Sanfey, 2019). On the other hand, people prefer some degree of inequality in society (Norton and Ariely, 2011; Fiske and Norton, 2014) and seem to be completely tolerant towards inequality, when it is considered to be a result of a fair process, such as an outcome of one's efforts rather than randomness (Brockner, 2002; Baumard, Mascar, Chevallie, 2012; Tyler, 2011; Almås et al., 2010), suggesting that the discourse should focus on unfairness-aversion rather than inequality-aversion (Starmans, Sheskin and Bloom, 2017). Inherent aversiveness of inequality remains also controversial in the light of studies investigating the link between income inequality and well-being. Although studies focusing on western samples have

found a negative link between averaged reported well-being and income inequality (Alesina, Di Tella, and MacCulloch, 2004; Oishi, Kesebir and Diener, 2013; Verme, 2011; Powdthavee, Burkhauser, and De Neve, 2017; O'Connel, 2004), studies including a greater variety of countries either find no relation (Berg and Veenhoven, 2010; Zagorski et al., 2013), or even a positive one (Kelley, and Evans, 2017; Rözer and Kraaykamp, 2013; Sanfey and Teksoz, 2007; Haller and Hadler, 2006; Katic and Ingram, 2017), suggesting that the relationship between inequality and well-being remains unclear. The latter result is often discussed in the context of relative deprivation theory, according to which inequality can be considered a positive phenomenon when it signals opportunity (Hirschman, 1973; Durongkaveroj, 2018). The support for this claim comes from the fact that a positive relationship between well-being and inequality has been mostly observed for developing or transitioning countries, in which inequality might spur optimism about a personal situation in the future (Kelley and Evans, 2017; Sanfey and Teksoz, 2007).

## REWARD CONTRASTS, REWARD ADAPTION, AND REFERENCE POINTS

Although undeniably social preferences modulate the value of rewards presented in distribution of rewards, many studies suggest that the same rewards can be also valued differently just because we gained experience with some other rewards. One of the earliest studies on the effect of previous experience on evaluative judgments found that subjective odour pleasantness depended on the presentation of preceding stimuli (Beebe-Center, 1929). Specifically, pleasantness increased and decreased after the presentation of the least and most pleasant odours respectively. Subsequent studies have revealed that affective judgments incorporate the whole history of experiences and extend beyond mere contrast effects of two stimuli (Helson, 1964). These findings led to a formulation of adaptation-level theory, according to which

affective and perceptual judgments are based on comparisons of stimulus to an internal norm that is updated with every experience (Helson, 1964). This norm, or adaptation-level, was mathematically expressed as a mean of the logarithm of previously experienced stimuli.

The adaptation-level theory was very influential across various disciplines. It is echoed in two components of prospect's theory: the idea that value of rewards and punishments diminishes with increasing magnitude (based on the law of diminishing marginal utility also present in classical economic theory), and a definition of a dynamically changing reference point that determines what is perceived as a gain or a loss (Tversky and Kahneman, 1992). The existence of reference points in reward evaluations has been demonstrated in many areas of decision-making since (Wang and Johnson, 2012; Rigoli, Friston, and Dolan, 2016; Bavard et al., 2018). The adaptation-level theory and the concept of a reference-point gathered also some support in economic surveys investigating the influence of pay on satisfaction, which identified contrast of one's income with a mean pay of a comparison group as a good predictor of satisfaction about one's income (Clark and Oswald, 1996; Byrgen, 2004).

The notion of a dynamically updated mean value is also closely related to reward expectation in the reinforcement learning literature (Bavard et al., 2018). The main difference between the two is that the latter assumes an exponentially decaying trace of past experiences instead of a simple average. One important study in this tradition has shown that changes in momentary subjective well-being can be well described as a weighted sum of such expectations and deviations of rewards from these expectations, known as prediction-errors (Rutledge et al., 2014).

## REWARD NORMALIZATION

Despite its generality (Helson, 1964), adaptation-level theory cannot explain many of the findings on evaluations of stimuli presented

simultaneously. In its simplest form, the comparison of two rewards is equivalent to taking a difference of their magnitudes. In principle, any agent should choose the reward with a higher magnitude. Could adding a third reward to the context change how the difference between the first two is perceived? According to the independence axiom of decision theory (Luce, 1959), the probability of selecting one reward over another should be independent of the existence of irrelevant alternatives. Contrary to this rational assumption, it has been demonstrated that adding an inferior option often causes an indifference or even preference reversal between the original options (Huber et al., 1982; Soltani, De Martino and Camerer, 2012; Louie et al., 2013). One biologically plausible explanation of this effect is based on a phenomenon of divisive normalization, in which response of a neuron is divided by a sum of neighbouring neurons (Louie et al., 2013; Carandini and Heeger, 2011). Divisive normalization is ubiquitous in the brain and is believed to solve the problem of efficient coding - that is, the problem of how the firing of a neuron should vary in response to different inputs, to most efficiently represent the available stimuli, given biophysical constraints on the firing range (Carandini and Heeger, 2011). According to this model, adding a third option decreases the difference between the two original options because the neural responses representing them are divided by a greater sum of neural activity.

The model of divisive normalization assumes that all stimuli present in context contribute equally to the computation of reward value. An alternative theory of range-normalization proposed that values are normalized with respect to the two most extreme values, corresponding to minimal and maximal possible firing rates of a neuron (Padoa-Schioppa, 2009; Soltani, De Martino and Camerer, 2012; Rustichini et al., 2017). In this theory, the representation of intermediate values would scale proportionally to these two extreme points. Consequently, adding a third option of inferior value should bring the value that was previously at the bottom closer to the top value,

causing a possible indifference between the two. This effect should be proportional to the distance between lower and intermediate value – a prediction that was supported by data (Soltani, De Martino and Camerer, 2012).

The above theories represent just a few examples of value-normalization processes (for a standard deviation normalization see: Diederen et al., 2016), that can be understood in a broad-terms as a transformation that forces two variables to be represented on the same scale. Normalization is also ubiquitous across different statistical and cognitive models that require scale-independence, including soft-max function widely used to model decision probability (Reverdy and Leonard, 2015) and Gini-coefficient used to quantify inequality (Lerman and Yitzhaki, 1984).

## REWARD RANK

Many of the above theories require encoding of the value of all stimuli and comparison of their magnitudes in absolute terms. However, research in psychophysics has established that humans are notoriously bad at making judgments about absolute values of stimuli, despite being quite good at discriminating them (Kinchla, 1971; Lockhead, 2004; Goffin and Olson, 2011). This led some researchers to suggest that judgments might rely not on absolute but ordinal comparisons (Parducci, 1992; Stewart, Brown, Chater, 2005; Stewart et al., 2006). Following this line of thought, the decision by sampling theory assumes that instead of comparing all values, we sample examples from past and present contexts and make affective judgments based on a value's rank in an ordered set of examples (Stewart et al., 2006). This theory was successful in recreating some of the classical effects of prospect theory, under the assumption that everyday distributions of probabilities, rewards and losses are consistent with a power-law function. The importance of rank has been also demonstrated in well-being studies, in which rank of pay, but not absolute pay, was shown to correlate with life-evaluations (Boyce, et

al., 2010) – suggesting that psychophysical comparisons are not the only domain in which ordinal comparisons might dominate judgments.

## INTERNAL SOCIAL CONTEXT

The above sections include some examples that could be considered an internal rather than external reward context, such as the influence of rewards experienced in the past. Another way in which internal context can change how reward signals are represented is through the interpretation of rewards based on personal beliefs and motivations. For example, it has been shown that people exhibit biased learning either after hearing bad news or hearing good news about the severity of climate change, depending on them being either climate change deniers or climate change believers (Sunstein et al., 2016). Therefore, in some cases, bad information can be favoured, if it confirms one's initial beliefs.

This topic has been also explored in research focused on preferences about social feedback. Two types of motivations have been proposed to regulate such preferences: self-enhancement and self-verification (Leary, 2007; Blaine and Crocker, 1993). According to the former, people are motivated to hold the best view of themselves and therefore will always prefer information that enhances their self-esteem. According to the later, people are motivated to hold consistent views of themselves, and therefore will seek feedback that confirms their beliefs. Depending on which motivation dominates, the same social feedback can be interpreted as more positive or more negative, irrespective of its absolute value (Kwang and Swann, 2010).

Internal social processes and motivations can also regulate the units of social comparisons. According to social identity theory, people not only possess personal identity, but also many group identities, such as their nationality or occupation (Ellemers, Knippenberg and Wilke, 1990). Therefore, in some situations, people might compare the relative standing of their group rather than their personal situation and react to it accordingly. For example, it

is known that people often derive a sense of pride from membership to the high-status group and incorporate it into their self-esteem (Smith and Tyler, 1997).

## MEASURING AFFECTIVE STATES

The last issue discussed concerns the measurement of affective state. Throughout the thesis, I rely on self-reported measures of well-being, which have been used in many previous studies before (Will et al., 2017; Rutledge et al., 2014; Rutledge et al., 2016; OECD, 2013). A common concern relates to the validity of such measures. However, it has been demonstrated that well-being self-reports correlate with external ratings of happiness of a person by family members (Zou, Schimmack, and Gere, 2013), facial expressions (Ito and Caccioppo, 1999), health indicators (Blanchflower and Oswald, 2007), suicide rates (Koivumaa-Honkanen et al., 2001), and self-reports on other related constructs (Sandvik, Diener, Seidlitz, 1993; Diener and Suh, 1997), suggesting that we can have relatively high confidence that such measures actually assess well-being.

## SUMMARY

Previous studies have demonstrated that the value of rewards is most likely constructed as a linear combination of the value derived from the properties the offered reward, such as its magnitude, and the value derived from contextual cues, such as the magnitudes of alternative rewards (Burke et al., 2016). Despite a plethora of evidence suggesting that context can change how rewards are evaluated, we still know little about the exact mechanisms through which context influences evaluation of rewards. In particular, in situations where the context consists of many alternative rewards, such as in the case of income distribution in the society, there are many possibilities of how context could transform the value of rewards. For example, the value of available reward could be normalized by the sum (Louie et al., 2013; Carandini

and Heeger, 2011) or range of all rewards in the context (Padoa-Schioppa, 2009; Soltani, De Martino and Camerer, 2012; Rustichini et al., 2017). Alternatively, it could be contrasted with the mean (Wang and Johnson, 2012; Rigoli, Friston, and Dolan, 2016; Bavard et al., 2018), or highest/lowest reward (Powdthavee et al., 2017). The absolute value of the available reward could be also ignored altogether, and instead expressed in relative terms as a rank of the reward in the distribution (Stewart, Brown, Chater, 2005; Stewart et al., 2006). Independently, preferences for specific distributions of rewards in the context could be also incorporated into the evaluation of the available rewards, as in the case of (un)fair distributions (Rutledge et al., 2016). Finally, it is also possible that the evaluation of rewards is influenced by internal contexts of personal motivations and higher order values (Sunstein et al., 2016; Leary, 2007; Blaine and Crocker, 1993).

## REFERENCES


Alesina, A., Di Tella, R., & MacCulloch, R. (2004). Inequality and happiness: Are Europeans and Americans different? *Journal of Public Economics*, *88*(9), 2009–2042. https://doi.org/10.1016/j.jpubeco.2003.07.006

Almås, I., Cappelen, A. W., Sørensen, E. Ø., & Tungodden, B. (2010). Fairness and the Development of Inequality Acceptance. *Science*, *328*(5982), 1176–1178. https://doi.org/10.1126/science.1187300

Baar, J. M. van, Chang, L. J., & Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature Communications*, *10*(1), 1–14. https://doi.org/10.1038/s41467-019-09161-6

Bardsley, N. (2008). Dictator game giving: Altruism or artefact? *Experimental Economics*, *11*(2), 122–133. https://doi.org/10.1007/s10683-007-9172-2

Battigalli, P., & Dufwenberg, M. (2007). Guilt in Games. *The American Economic Review*, *97*(2), 170–176. JSTOR.

Baumard, N., Mascaro, O., & Chevallier, C. (2012). Preschoolers are able to take merit into account when distributing goods. *Developmental Psychology*, *48*(2), 492–498. https://doi.org/10.1037/a0026598

Bavard, S., Lebreton, M., Khamassi, M., Coricelli, G., & Palminteri, S. (2018). Reference-point centering and range-adaptation enhance human

reinforcement learning at the cost of irrational preferences. *Nature Communications*, *9*(1), 1–12. https://doi.org/10.1038/s41467-018-06781-2

Bechtel, M. M., Liesch, R., & Scheve, K. F. (2018). Inequality and redistribution behavior in a give-or-take game. *Proceedings of the National Academy of Sciences*, *115*(14), 3611–3616. https://doi.org/10.1073/pnas.1720457115

Beebe-Center, J. G. (1929). The Law of Affective Equilibrium. *The American Journal of Psychology*, *41*(1), 54–69. JSTOR. https://doi.org/10.2307/1415108

Berg, M., & Veenhoven, R. (2010). *Income inequality and happiness in 119 nations*. https://doi.org/10.4337/9781781000731.00017

Blaine, B., & Crocker, J. (1993). Self-Esteem and Self-Serving Biases in Reactions to Positive and Negative Events: An Integrative Review. In R. F. Baumeister (Ed.), *Self-Esteem: The Puzzle of Low Self-Regard* (pp. 55–85). Springer US. https://doi.org/10.1007/978-1-4684-8956-9_4

Blanchflower, D. G., & Oswald, A. J. (2008). Hypertension and happiness across nations. *Journal of Health Economics*, *27*(2), 218–233. https://doi.org/10.1016/j.jhealeco.2007.06.002

Boyce, C. J., Brown, G. D. A., & Moore, S. C. (2010). Money and Happiness: Rank of Income, Not Income, Affects Life Satisfaction. *Psychological Science*, *21*(4), 471–475. JSTOR.

Brockner, J. (2002). Making Sense of Procedural Fairness: How High Procedural Fairness Can Reduce or Heighten the Influence of Outcome Favorability. *Academy of Management Review*, *27*(1), 58–76. https://doi.org/10.5465/amr.2002.5922363

Burke, C. J., Baddeley, M., Tobler, P. N., and Schultz, W. (2016). Partial Adaptation of Obtained and Observed Value Signals Preserves Information about Gains and Losses. *The Journal of Neuroscience*, 36(39), 10016–10025. https://doi.org/10.1523/JNEUROSCI.0487-16.2016

Bygren, M. (2004). Pay reference standards and pay satisfaction: What do workers evaluate their pay against? *Social Science Research*, *33*(2), 206–224. https://doi.org/10.1016/S0049-089X(03)00045-0

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*(1), 51–62. https://doi.org/10.1038/nrn3136

Clark, A. E., & Oswald, A. J. (1996). Satisfaction and comparison income. *Journal of Public Economics*, *61*(3), 359–381. https://doi.org/10.1016/0047-2727(95)01564-7

Diederen, K. M. J., Spencer, T., Vestergaard, M. D., Fletcher, P. C., & Schultz, W. (2016). Adaptive Prediction Error Coding in the Human Midbrain and Striatum Facilitates Behavioral Adaptation and Learning Efficiency. *Neuron*, *90*(5), 1127–1138. https://doi.org/10.1016/j.neuron.2016.04.019

Diener, E., & Suh, E. (1997). MEASURING QUALITY OF LIFE: ECONOMIC, SOCIAL, AND SUBJECTIVE INDICATORS. *Social Indicators Research*, *40*(1), 189–216. https://doi.org/10.1023/A:1006859511756

Durongkaveroj, W. (2018). Tolerance for inequality: Hirschman's tunnel effect revisited. *Journal of International Development*, *30*(7), 1240–1247. https://doi.org/10.1002/jid.3389

Eagleman, D. M. (2001). Visual illusions and neurobiology. *Nature Reviews Neuroscience*, *2*(12), 920–926. https://doi.org/10.1038/35104092

Ehrenfels, C. von. (1890). *Über 'Gestaltqualitäten'*. Reisland.

Ellemers, N., van Knippenberg, A., & Wilke, H. (1990). The influence of permeability of group boundaries and stability of group status on strategies of individual mobility and social change. *The British Journal of Social Psychology*, *29 ( Pt 3)*, 233–246. https://doi.org/10.1111/j.2044-8309.1990.tb00902.x

Engel, C. (2011). Dictator games: A meta study. *Experimental Economics*, *14*(4), 583–610. https://doi.org/10.1007/s10683-011-9283-7

Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. https://doi.org/10.1162/003355399556151

Goffin, R. D., & Olson, J. M. (2011). Is It All Relative? Comparative Judgments and the Possible Improvement of Self-Ratings and Ratings of Others. *Perspectives on Psychological Science*, *6*(1), 48–60. JSTOR.

Haller, M., & Hadler, M. (2006). How Social Relations and Structures can Produce Happiness and Unhappiness: An International Comparative Analysis. *Social Indicators Research*, *75*(2), 169–216. https://doi.org/10.1007/s11205-004-6297-y

Helson, H. (1964). *Adaptation-level theory: An experimental and systematic approach to behavior*. New York.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies. *The American Economic Review*, *91*(2), 73–78. JSTOR.

Hirschman, A. O. (1973). The changing tolerance for income inequality in the course of economic development. *World Development*, *1*(12), 29–36. https://doi.org/10.1016/0305-750X(73)90109-5

Huber, J., Payne, J. W., & Puto, C. (1982). Adding Asymmetrically Dominated Alternatives: Violations of Regularity and the Similarity Hypothesis. *Journal of Consumer Research*, *9*(1), 90–98. JSTOR.

Ito, T. A., & Cacioppo, J. T. (1999). The psychophysiology of utility appraisals. In *Well-being: The foundations of hedonic psychology* (pp. 470–488). Russell Sage Foundation.

Jäkel, F., Singh, M., Wichmann, F. A., & Herzog, M. H. (2016). An overview of quantitative approaches in Gestalt perception. *Vision Research*, *126*, 3–8. https://doi.org/10.1016/j.visres.2016.06.004

Katic, I., & Ingram, P. (2018). Income Inequality and Subjective Well-Being: Toward an Understanding of the Relationship and Its Mechanisms. *Business & Society*, *57*(6), 1010–1044. https://doi.org/10.1177/0007650317701226

Kelley, J., & Evans, M. D. R. (2017). Societal Inequality and individual subjective well-being: Results from 68 societies and over 200,000 individuals, 1981-2008. *Social Science Research*, *62*, 1–23. https://doi.org/10.1016/j.ssresearch.2016.04.020

Kinchla, R. A. (1971). Visual movement perception: A comparison of absolute and relative movement discrimination. *Perception & Psychophysics*, *9*(2), 165–171. https://doi.org/10.3758/BF03212622

Koivumaa-Honkanen, H., Honkanen, R., Viinamäki, H., Heikkilä, K., Kaprio, J., & Koskenvuo, M. (2001). Life satisfaction and suicide: A 20-year follow-up study. *The American Journal of Psychiatry*, *158*(3), 433–439. https://doi.org/10.1176/appi.ajp.158.3.433

Kwang, T., & William B. Swann, J. (2010). Do People Embrace Praise Even When They Feel Unworthy? A Review of Critical Tests of Self-Enhancement Versus Self-Verification: *Personality and Social Psychology Review*. https://doi.org/10.1177/1088868310365876

Lerman, R. I., & Yitzhaki, S. (1984). A note on the calculation and interpretation of the Gini index. *Economics Letters*, *15*(3), 363–368. https://doi.org/10.1016/0165-1765(84)90126-5

List, J. A. (2007). On the Interpretation of Giving in Dictator Games. *Journal of Political Economy*, *115*(3), 482–493. JSTOR. https://doi.org/10.1086/519249

Lockhead, G. R. (2004). Absolute Judgments Are Relative: A Reinterpretation of Some Psychophysical Ideas. *Review of General Psychology*, *8*(4), 265–272. https://doi.org/10.1037/1089-2680.8.4.265

Louie, K., Khaw, M. W., & Glimcher, P. W. (2013). Normalization is a general neural mechanism for context-dependent decision making. *Proceedings of the National Academy of Sciences*, *110*(15), 6139–6144. https://doi.org/10.1073/pnas.1217854110

Luce, R. D. (2005). *Individual choice behavior: A theoretical analysis*. Dover Publications. https://doi.org/10.1037/14396-000

Nihonsugi, T., Ihara, A., & Haruno, M. (2015). Selective Increase of Intention-Based Economic Decisions by Noninvasive Brain Stimulation to the Dorsolateral Prefrontal Cortex. *Journal of Neuroscience*, *35*(8), 3412–3419. https://doi.org/10.1523/JNEUROSCI.3885-14.2015

Norton, M. I. (2014). Unequality: Who Gets What and Why It Matters. *Policy Insights from the Behavioral and Brain Sciences*, *1*(1), 151–155. https://doi.org/10.1177/2372732214550167

Norton, M. I., & Ariely, D. (2011). Building a Better America—One Wealth Quintile at a Time. *Perspectives on Psychological Science*, *6*(1), 9–12. https://doi.org/10.1177/1745691610393524

O'Connell, M. (2004). Fairly satisfied: Economic equality, wealth and satisfaction. *Journal of Economic Psychology*, *25*(3), 297–305. https://doi.org/10.1016/S0167-4870(03)00010-2

OECD. (2013). *OECD Guidelines on Measuring Subjective Well-being*. OECD Publishing.

Oishi, S., Kesebir, S., & Diener, E. (2011). Income Inequality and Happiness. *Psychological Science*, *22*(9), 1095–1100. https://doi.org/10.1177/0956797611417262

Padoa-Schioppa, C. (2009). Range-Adapting Representation of Economic Value in the Orbitofrontal Cortex. *Journal of Neuroscience*, *29*(44), 14004–14014. https://doi.org/10.1523/JNEUROSCI.3751-09.2009

Parducci, A. (1995). *Happiness, pleasure, and judgment: The contextual theory and its applications*. Lawrence Erlbaum Associates, Inc.

Powdthavee, N., Burkhauser, R. V., & De Neve, J.-E. (2017). Top incomes and human well-being: Evidence from the Gallup World Poll. *Journal of Economic Psychology*, *62*, 246–257. https://doi.org/10.1016/j.joep.2017.07.006

Reverdy, P., & Leonard, N. E. (2016). Parameter Estimation in Softmax Decision-Making Models With Linear Objective Functions. *IEEE*

*Transactions on Automation Science and Engineering*, *13*(1), 54–67. https://doi.org/10.1109/TASE.2015.2499244

Rigoli, F., Friston, K. J., & Dolan, R. J. (2016). Neural processes mediating contextual influences on human choice behaviour. *Nature Communications*, *7*(1), 1–11. https://doi.org/10.1038/ncomms12416

Rözer, J., & Kraaykamp, G. (2013). Income Inequality and Subjective Well-being: A Cross-National Study on the Conditional Effects of Individual and National Characteristics. *Social Indicators Research*, *113*(3), 1009–1023. https://doi.org/10.1007/s11205-012-0124-7

Rustichini, A., Conen, K. E., Cai, X., & Padoa-Schioppa, C. (2017). Optimal coding and neuronal adaptation in economic decisions. *Nature Communications*, *8*(1), 1–14. https://doi.org/10.1038/s41467-017-01373-y

Rutledge, R. B., Berker, A. O. de, Espenhahn, S., Dayan, P., & Dolan, R. J. (2016). The social contingency of momentary subjective well-being. *Nature Communications*, *7*(1), 1–8. https://doi.org/10.1038/ncomms11825

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, *111*(33), 12252–12257. https://doi.org/10.1073/pnas.1407535111

Sandvik, E., Diener, E., & Seidlitz, L. (2009). Subjective Well-Being: The Convergence and Stability of Self-Report and Non-Self-Report Measures. In E. Diener (Ed.), *Assessing Well-Being: The Collected Works of Ed Diener* (pp. 119–138). Springer Netherlands. https://doi.org/10.1007/978-90-481-2354-4_6

Sanfey, P., & Teksoz, U. (2007). Does transition make you happy?1. *Economics of Transition and Institutional Change*, *15*(4), 707–731. https://doi.org/10.1111/j.1468-0351.2007.00309.x

Smith, H. J., & Tyler, T. R. (1997). Choosing the Right Pond: The Impact of Group Membership on Self-Esteem and Group-Oriented Behavior. *Journal of Experimental Social Psychology*, *33*(2), 146–170. https://doi.org/10.1006/jesp.1996.1318

Soltani, A., Martino, B. D., & Camerer, C. (2012). A Range-Normalization Model of Context-Dependent Choice: A New Model and Evidence. *PLOS Computational Biology*, *8*(7), e1002607. https://doi.org/10.1371/journal.pcbi.1002607

Starmans, C., Sheskin, M., & Bloom, P. (2017). Why people prefer unequal societies. *Nature Human Behaviour*, *1*(4), 1–7. https://doi.org/10.1038/s41562-017-0082

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review*, *112*(4), 881–911. https://doi.org/10.1037/0033-295X.112.4.881

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26. https://doi.org/10.1016/j.cogpsych.2005.10.003

Sunstein, C. R., Bobadilla-Suarez, S., Lazzaro, S. C., & Sharot, T. (2016). How People Update Beliefs about Climate Change: Good News and Bad News. *Cornell Law Review*, *102*, 1431.

Tricomi, E., Rangel, A., Camerer, C. F., & O'Doherty, J. P. (2010). Neural evidence for inequality-averse social preferences. *Nature*, *463*(7284), 1089–1091. https://doi.org/10.1038/nature08785

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. https://doi.org/10.1007/BF00122574

Tyler, T. (2011). Procedural Justice Shapes Evaluations of Income Inequality: Commentary on Norton and Ariely (2011). *Perspectives on Psychological Science*, *6*(1), 15–16. https://doi.org/10.1177/1745691610393981

Verme, P. (2011). Life Satisfaction and Income Inequality. *Review of Income and Wealth*, *57*(1), 111–127. https://doi.org/10.1111/j.1475-4991.2010.00420.x

Waal, F. B. M. de, & Waal, F. de. (1996). *Good Natured*. Harvard University Press.

Wang, X. T., & Johnson, J. G. (2012). A tri-reference point theory of decision making under risk. *Journal of Experimental Psychology. General*, *141*(4), 743–756. https://doi.org/10.1037/a0027415

Zagorski, K., Evans, M. D. R., Kelley, J., & Piotrowska, K. (2014). Does National Income Inequality Affect Individuals' Quality of Life in Europe? Inequality, Happiness, Finances, and Health. *Social Indicators Research*, *117*(3), 1089–1110. JSTOR.

Zou, C., Schimmack, U., & Gere, J. (2013). The validity of well-being measures: A multiple-indicator-multiple-rater model. *Psychological Assessment*, *25*(4), 1247–1254. https://doi.org/10.1037/a0033902

# Chapter 2

# The Motivational Cost of Inequality: Opportunity Gaps Reduce The Willingness To Pursue Rewards

**Filip Gesiarz[*1], Jan-Emmanuel De Neve[2], Tali Sharot[*1]**

[1]Affective Brain Lab, Department of Experimental Psychology, University College London, London, UK

[2]Saïd Business School, University of Oxford, Oxford, UK

## ABSTRACT

Factors beyond a person's control, such as demographic characteristics at birth, often influence the availability of rewards an individual can expect for their efforts. We know surprisingly little how such differences in opportunities impact human motivation. To test this, we designed a study in which we arbitrarily varied the reward offered to each participant in a group for performing the same task. Participants then had to decide whether or not they were willing to exert effort to receive their reward. Across three experiments, we found that the unequal distribution of offers reduced participants' motivation to pursue rewards even when their relative position in the distribution was high, and despite the decision being of no benefit to others and reducing the reward for oneself. Participants' feelings partially mediated this relationship. In particular, a large disparity in rewards was associated with greater unhappiness, which was associated with lower willingness to work – even when controlling for absolute reward and its relative value, both of which also affected decisions to work. A model that incorporated a person's relative position and unfairness of rewards in the group fit better to the data than other popular models describing the effects of inequality. Our findings suggest opportunity-gaps can trigger psychological dynamics that hurt productivity and well-being of all involved.

## INTRODUCTION

Randomness plays a surprisingly important role in determining the barriers and opportunities encountered by individuals on their path to a prosperous life (Pluchino et al., 2018). Country of birth alone explains 66% of global variation in living standards (Milanovic, 2014). Other non-meritocratic factors, such as zip code (Chetty & Hendren, 2018), parental socio-economic status (Duncan & Magnuson, 2012), gender (Blau & Kahn, 2007), or a person's name (Silberzahn & Uhlmann, 2013) have been shown to have a significant effect on earnings, even after accounting for inter-individual differences in merit. Economic inequality arising due to random circumstances is often viewed as unfair (Starmans et al., 2017), and previous studies have shown that people support redistribution of wealth in such situations (McCall et al., 2017). However, much less is known about how opportunity gaps influence human motivation. Such knowledge could shed light on psychological mechanisms that lead to differences in aspirations, that in turn might contribute to higher unemployment (Elmelech & Lu, 2004; Findlay & Wright, 1996; Uhrig, 2015) and lower university application rates of people from disadvantaged backgrounds (Boliver, 2013; Crawford et al., 2014; Thiele et al., 2017). Here, we examine how randomly assigned unequal reward prospects can influence a person's willingness to exert effort in exchange for rewards – a proxy measure of motivation in labour supply decisions.

Due to a lack of experimental research on the impact of inequality on motivation, the underlying mechanisms of this relationship remain unknown. We hypothesize that arbitrary differences in opportunities to earn rewards can negatively impact not only disadvantaged individuals but also those who are offered relatively high rewards. This is because facing opportunity gaps can involve two separate mechanisms: relative comparisons and reactions to unfairness, representing self-regarding and group-regarding reactions to inequality, respectively (Clark & D'Ambrosio, 2015). First, because people engage in spontaneous social comparisons, evaluating their rewards relative

to those of others (Bault et al., 2011; Boyce et al., 2010; Hagerty, 2000; Lyubomirsky & Ross, 1997), opportunity gaps can increase motivation to pursue rewards of those offered relatively high rewards and reduce the motivation of those offered relatively low rewards. However, at the same time people may have a negative response to the unfairness of arbitrary distributions of rewards in their group regardless of which side of the distribution they are at, and be less willing to pursue rewards in situations that are unfair. Indeed, it has been shown that subjects are less happy when they themselves win in a gambling task, but the other subject loses, in comparison to when both subjects win (Rutledge, de Berker, et al., 2016). We hypothesize that such a negative reaction may have consequences beyond a person's affective state. Specifically, negative feelings can lead to apathy as well as a reduction in the subjective value of rewards (Eldar & Niv, 2015), leading to a reduced motivation of all members of the group. Thus, individuals at the bottom of the distribution may be negatively affected twice, first due to their lower relative position and second due to their reaction to unfair distribution.

We formalize the above hypotheses in a model that characterizes the motivational response to rewards as a linear combination of reward's absolute value, relative value, and statistical dispersion of all rewards in the group. Based on the law of decreasing marginal utility, we assume that absolute reward has a non-linear effect on decisions to engage in an effort to earn the reward (Tversky & Kahneman, 1992). As previous studies have shown that people have a tendency to engage in ordinal rather than absolute comparisons (Stewart et al., 2006), we define the relative value of rewards as the rank of the offered reward. Statistical dispersion is calculated in our model as Gini coefficient, following other studies suggesting a relation between this measure and well-being in national surveys (Oishi et al., 2011).

In three experiments, we were able to dissociate and quantify the influence unfairness, reward's rank, its absolute value, while studying them independently from other factors that are often associated with opportunity

gaps, such as demographics or stereotypes. In all three studies, participants made decisions on whether to exert cognitive effort in exchange for a reward while observing the rewards offered to others for completing the same task. In these experiments we manipulated: (i) the deviation of payments in the group from an equal distribution (thereafter 'unfairness'), and (ii) the relative position of the offer in the distribution (thereafter 'rank'). Experiment 1 aimed to establish if the motivation to work for rewards is influenced by unfairness and rank of offered rewards. Experiments 2 and 3 aimed to test the mechanisms underlying the influence of relative value and unfairness on motivation, including the mediating role of emotions, and the moderating role of uncertainty.

## METHODS EXPERIMENT 1

### Overview

Experiment 1 used a one-shot design (Fig 1) and was conducted on Prolific - an online labour market platform. Participants were offered £0.24 for an optional task of transcribing 1/3 of a page of text from a displayed image and were made to believe that this reward offer was drawn at random. The offered reward was displayed in the context of 4 other rewards assigned randomly to other workers on the platform. Seven hundred participants were assigned to separate conditions in a 2x5 design that determined the context of their offered reward: the five rewards could be either relatively equally distributed or unequally distributed, and participant's reward of £0.24 could be presented either as $5^{th}$, $4^{th}$, $3^{rd}$, $2^{nd}$ or $1^{st}$ best-offered reward.


**Fig 1. Behavioural task in Experiment 1.** (A) The online task had a one-shot design. The task was advertised on an online labour market platform as a simple transcribing job. After completing the mandatory cognitive task, participants were informed that they would have an opportunity to complete an optional transcribing task for a randomly drawn fee. The random draw was determined by spinning a wheel of fortune that assigned a participant one out of five different colours. After the colour assignment, participants were presented with the reward offers for all five colours, and were told that the other rewards were assigned to other people who drew those colours. Participants then decided to either accept or decline the reward offer for the optional task. If they accepted it, they had to transcribe an additional text. If they declined it, the task ended, and they were granted their base fee. Unbeknownst to participants, the reward offer was always equal to £0.24, and the random colour assignment determined if the offer was presented either as $5^{th}$, $4^{th}$, $3^{rd}$, $2^{nd}$ or $1^{st}$ best reward. Independently, participants were randomly assigned to one of two levels of unfairness (Gini coefficient = 0.3 or Gini coefficient = 0.5) of reward distribution. (B) The task had a 5x2 design (10 conditions in total): two levels of inequality, and five levels of relative value. Each participant viewed only one of these conditions. Panel B illustrates example conditions. The example a) shows a situation

where £0.24 was presented as the second-best reward in an unfair context, and example b) where it was presented as the second-best reward in a fair context. The example c) shows a situation where £0.24 was presented as the middle reward in an unfair context, and the example d) shows a situation where £0.24 was presented as the middle reward in a fair context.

## Participants

In experiment 1, seven hundred participants were recruited to take part in the online study, spread evenly across ten conditions (70 participants per condition). All participants provided written informed consent. The experiment was approved by the UCL ethics committee. Participants were recruited through the Prolific platform – an online platform for offering web-based tasks. Eighty participants were excluded due to failing attention check that asked them about the colour that they have been assigned to. This exclusion criterion was necessary, as the colour indicated which reward a participant was offered. All participants in the online task were currently UK residents (mean age 26.2[5.0], age range 18 – 35, 487 women). The average self-identified political orientation was 4.61(1.61) on a scale ranging from 1 (extremely right-wing) to 7 (extremely left-wing), significantly more left-wing than the centre of the scale ($t(699) = 18.27$, $p < 0.001$).

## Procedure

Participants responded to an ad on Prolific platform that recruited people for a short transcribing task – a common task on online labour markets. The display of the advertisement was restricted to current UK residents aged 18 – 35. After signing up to complete the task, participants were informed that the task will consist of a mandatory transcribing task, for which they will be paid the advertised wage (£0.25), and an optional transcribing task for which they will be paid a bonus payment. The mandatory transcribing task required participants to transcribe a 1/5 of a page from an old cookbook. The optional

task was to transcribe a different text from the same cookbook, which was approximately 3 times longer. The instructions emphasized that participants had to be 99% accurate to receive the bonus payment. There was no time limit.

They were also informed that the wage for the optional task would be randomly drawn. The random draw was determined by a wheel of fortune that after spinning for 3 seconds picked one colour out of 5 colours. After the participant was assigned one of 5 colours, the bonus wages for the optional task were revealed all at once for all 5 colours. Participants were told that information about the other wages was displayed to inform other Prolific users who drew different colours. Unbeknownst to participants, the offered wage for the optional task was always equal to £0.24, and each participant was assigned to one condition in a 2x5 design that determined the context in which the reward was displayed. In particular, the reward could be presented either as $5^{th}$, $4^{th}$, $3^{rd}$, $2^{nd}$ or $1^{st}$ best reward, and was presented either in a context of a roughly fair distribution of rewards between participants (corresponding to a 0.2 Gini coefficient) or an unfair distribution (corresponding to a 0.5 Gini coefficient). Full list of reward distributions is included in the Supporting Table 2. After seeing the reward offers, participants had to decide to either accept or reject the optional task. If they decided to accept it, they had to transcribe an additional text and were paid their bonus wage (£0.24) plus base wage (£0.25). If they decided to reject it, they were paid just their base wage (£0.25).

**Data analysis**

To test the influence of reward's rank and unfairness of the distribution, we used a Generalized Linear Model (GLM) that included decisions to work as the categorical dependent variable, and unfairness (measured as Gini coefficient) and rank (normalized to range from 0 to 1, for lowest and highest rank respectively) as independent variables. Both independent variables were standardized prior to the analysis. The GLME model assumed a binomial distribution of the dependent variable.

Participant's offer rank was normalized to range from 0 to 1 as follows:

$$Rank_t = \frac{i-1}{n-1}$$

Where *i* is the reward offer index in a set of offers ordered from lowest to highest and *n* is the number of participants in the group (in our case 5). The above rank measure assigns 1 to the person with the best offer, 0 to the person with the lowest offer, and 0.5 to the person with the intermediate offer. Unfairness was measured as the Gini coefficient, calculated as follows:

$$Unfairness = \frac{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2\bar{x}}$$

Where *n* is the number of participants in the group, $x_i$ and $x_j$ is the reward offers received by each person, and $\bar{x}$ is the mean reward offer.

To illustrate the results from experiment 1, we plotted the number of participants who decided to pursue additional reward divided by the number of all participants in the condition separately for each rank and level of unfairness (Fig 2).

## RESULTS EXPERIMENT 1

Overall, 77.33% of participants decided to perform the optional task in exchange for an additional fee of £0.24. However, we found that participants were less willing to work for the additional reward when they believed that the distributions of offered rewards were unfair vs. fair ($\beta = -0.31$, $p < 0.01$), and when the rank of their reward was low vs. high ($\beta = 0.40$, $p < 0.001$), despite absolute reward being the same across all conditions in this experiment (Fig 2). On average, an increase of 0.3 in Gini coefficient resulted in 10.6% less

accepted offers, and increase of one rank resulted in 5.3% more accepted offers.

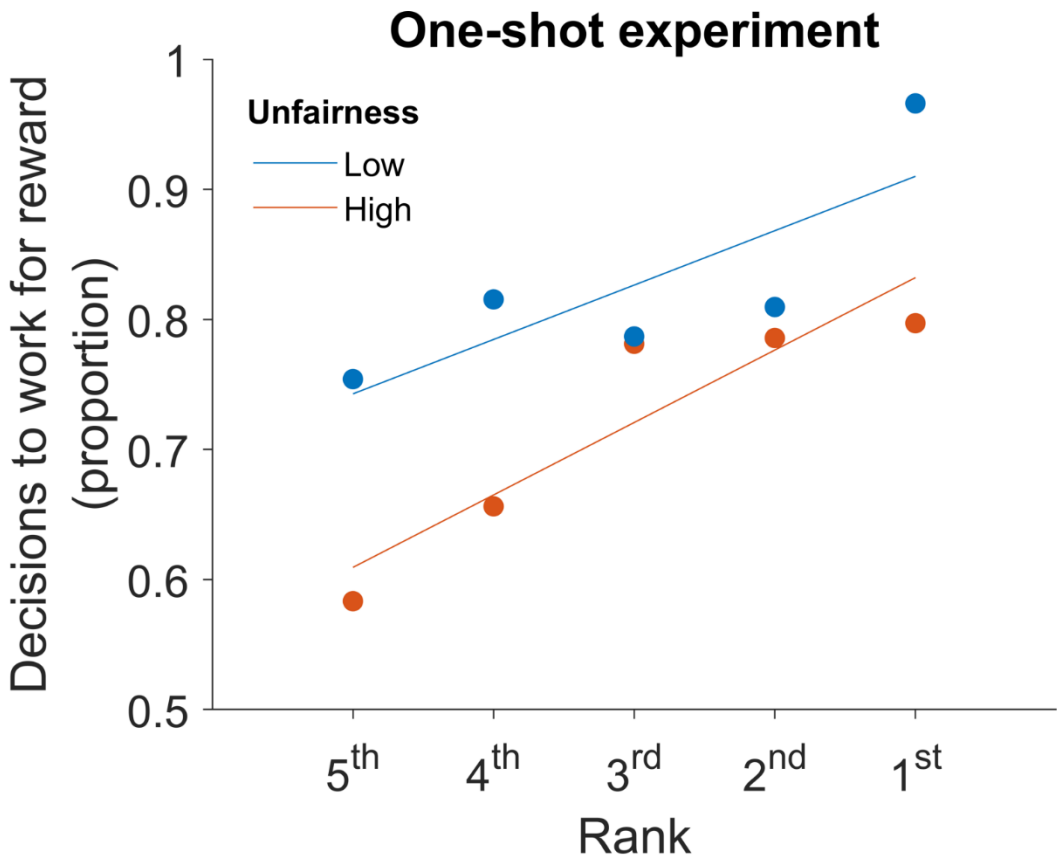


**Fig 2**. **Motivation to work is higher when the distribution of rewards is fair and the rank of the reward is high, despite the same level of absolute reward**. The plot illustrates the results from a one-shot experiment conducted on an online labour market platform. Each dot represents the proportion of participants who decided to perform an additional task for a bonus reward of £0.24, which was presented either in a relatively fair (blue) or unfair (red) context, and either as the $5^{th}$, $4^{th}$, $3^{rd}$, $2^{nd}$ or $1^{st}$ best reward. The lines represent the best fitting line based on the Ordinary Least Squares method. Participants were more likely to accept the offer of £0.24 when its rank was high than when it was low, and when the rewards of all participants were fairly distributed than when they were unfairly distributed.

## METHODS EXPERIMENTS 2 & 3

### Overview

Experiments 2 and 3 followed a similar logic as experiment 1, but used a repeated measure design (Fig 3), in which the same person was exposed to different distributions of rewards. Repeated measure designs achieve greater statistical power with fewer participants, allowing us to test more efficiently a larger number of hypotheses regarding the mechanisms underlying the effects observed in Experiment 1. Experiments 2 and 3 aimed to test the robustness of the effects observed in Experiment 1 when translated to a different context. Both experiments included a greater variety of distributions (both positively skewed and negatively skewed) and a different cognitive effort task. Different distributions allowed us to test the predictions of different models describing the impact of inequality on the evaluation of rewards. Additionally, the experiments: a) gathered information about participants' current feelings after seeing the distribution of rewards, allowing us to test if the observed effects are mediated by the impact of rank and unfairness on person's emotional state, and b) manipulated the uncertainty about the value of rewards, by either introducing a known (Experiment 2) or an unknown (Experiment 3) exchange rate of earned points with £, allowing us to test if reliance on the social context in one's decisions to work is moderated by uncertainty.

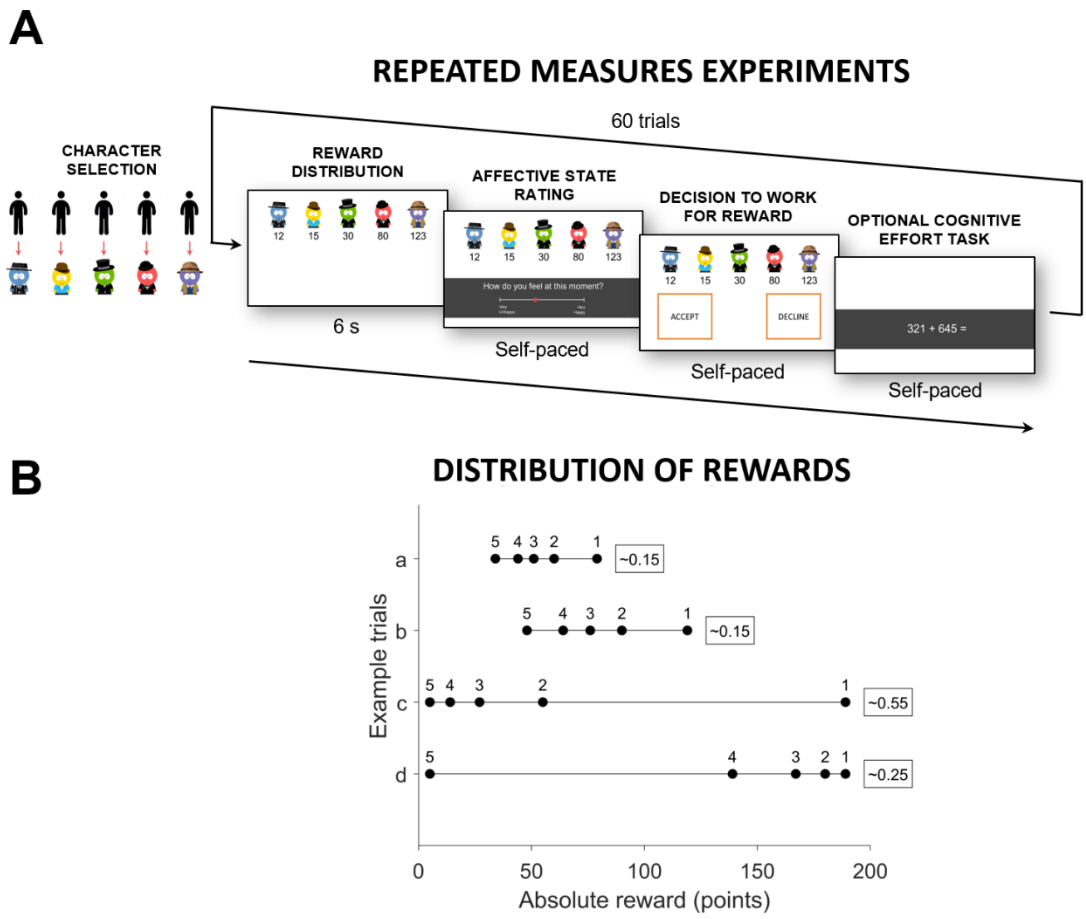


**Fig 3. Behavioural task in Experiment 2 and 3.** (A) Both Experiment 2 and 3 used a repeated measures design. Participants were invited to the lab in groups of five. To easily identify themselves during the experiment, each participant selected a cartoon avatar that would represent them throughout the task. They then retired to individual cubicles to complete the study. There were 60 trials in total. Each trial started with a display of all participants' reward offers, that differed in rank, absolute reward, and level of unfairness between participants. After seeing the distribution of rewards, participants rated their current feelings and indicated whether they were willing to exert cognitive effort for their offered reward on that specific trial. If they decided to do so, they would complete three mathematical problems. If not, they would move on to the next trial. If a participant gave an incorrect answer to the mathematical problem, they would have to solve an additional one, until they completed three problems correctly. (B) For the repeated measure experiments, we created 30 income distributions based on a log-normal probability density function (corresponding to 10 levels of Gini index uniformly distributed between 15 and 55, with 3 different median values) Log-normal distribution approximates reward distributions encountered in real-world, such as income distributions within countries (Pinkovskiy & Sala-i-Martin, 2009) and companies (Lazear & Shaw, 2009). Because these distributions are always positively skewed, we also created 30

distributions that were negatively skewed and a mirror image of the positively skewed distributions. For illustration purposes, we plot four of these distributions. Each dot on the line represents one of five reward offers presented to participants. Numbers above the dots refer to the reward's rank. Numbers in the rectangles refer to unfairness level, expressed in the Gini coefficient. Distribution a) is an example of a fair positively skewed distribution with a low median reward; distribution b) is an example of similarly fair distribution, but with a higher median value; Distribution c) is an example of an unfair positively skewed distribution, and distribution d) is an example of a negatively skewed distribution that is a mirror image of c).

**Participants**

In Experiment 2 and 3, one hundred and ten participants from University College London subject pool were recruited to take part in two onsite studies: sixty in experiment 2 (mean age 22.1[3.2], age range 18 – 35; 38 women) and fifty in experiment 3 (mean age 21.4[2.0]; age range 18 – 35; 34 women). All participants provided written informed consent. The experiment was approved by the UCL ethics committee. Across these two experiments, 67% of participants originated from Western countries. The average self-identified political orientation was 3.52(1.38) on a scale ranging from 1 (extremely right-wing) to 7 (extremely left-wing) and was not significantly different from the centre of the scale ($t(87)=0.12$, $p = 0.91$). All participants started with an initial endowment of £10 and were paid an additional bonus based on their decision to accept or reject reward offers in exchange for performing a cognitive task in one randomly selected trial. Participants who accepted all reward offers were excluded from the data analysis as we could not identify the factors influencing their decisions due to lack of behavioral variability, beyond the fact that they were maximizing their bonus reward at the end (eight subjects in experiment 2 and seven subjects in experiment 3), leaving 52 and 43 participants in each experimental sample respectively. None of the subjects rejected all offers.

**Task**

In both experiments, we invited participants to the lab in groups of five ($N = 110$ in total). To easily identify themselves during the task, participants

were asked to choose a cartoon avatar that would represent them in the study. A randomly drawn lot number determined the order of choosing avatars. Participants were informed that each person will be offered a different reward on each trial and that these rewards were randomly decided on each trial by a computer program. Next, participants retired to separate cubicles where they were given additional instructions.

Participants first completed one practice trial. Both experiments consisted of 60 trials. In each of 60 trials, we presented to participants the reward points offered to each of the five members of the group on that trial. On each trial, we independently manipulated: (i) the deviation of payments in the group from an equal distribution ('*unfairness*'), (ii) the rank of the reward offered to each person within the group (ranging from 1 to 5 - '*rank*') and (iii) the absolute reward offered (i.e., points - '*absolute reward*').

We created 60 different distributions of reward offers in total and presented them in random order. We generated 30 reward distributions based on a log-normal probability density function. Log-normal distribution was chosen as it fits closely real-world income structures within firms (Lazear & Shaw, 2009) and countries (Pinkovskiy & Sala-i-Martin, 2009). To vary the levels of reward magnitude range and statistical dispersion we used a combination of 3 different median values (0.55, 1, 1.45) and ten different standard deviations, corresponding to values of the Gini coefficient varying uniformly from 20 to 65, resulting in 30 different distributions. Log-normal distributions are always positively skewed. To generalize our findings, we also included 30 negatively skewed distributions that were a mirror-image of the positively skewed distributions by applying the following transformation of representative values:

$$x_{positive} = \{x_1, x_2, x_3, x_4, x_5\}$$

$$x_{negative} = \left|x_{positive} - max\left(x_{positive}\right)\right| + \min\left(x_{positive}\right)$$

Where $x_n$ is subject $n$ payment offer in each trial, $x_{positive}$ and $x_{negative}$ are payment offers of all participants in trials with positively and negatively skewed distributions, respectively.

To generate reward offers representative of the above distributions, we used an inverse cumulative density function of these distributions, which assigns maximal pay value earned by each percent of the population. We next took an average pay from subsequent 20 percentiles of this function, with the exclusion of top 1 percentile, resulting in 5 values reflecting an average pay of each 20% of the population. The last percentile was excluded as it approaches infinity. Unfairness was quantified based on these 5 representative values. To introduce variability to the middle pay (that otherwise would be the same for all distributions generated from the same median value) we additionally subtracted a number between 0 and 9 from each representative value in each distribution (in each distribution the same number was subtracted for each value). This resulted in the pay offers shown in Supporting Table 1.

After seeing the distribution of reward offers, participants then rated their feelings by clicking on a continuous sliding scale ranging from very unhappy to very happy. The slider started in the middle of the scale on every trial. After the feeling ratings, participants indicated whether they were willing to complete three mathematical problems to earn their reward. If they decided to do so, they were asked to solve the problems (the instructions emphasized that the mathematical problems were the same for all). If they decided not to, they would move on to the next trial. Each problem required adding two 3-digit numbers. To ensure equal difficulty of mathematical problems throughout the task, each addition had exactly two carryovers (sum of ones, tens or hundreds greater than 10). E.g., problems included sums like 118 + 197. If participants provided an incorrect answer, they had to solve an additional problem. Participants continued until they got three problems correct. On average, 89% of attempts were correct, and it took subjects 17 seconds (SD = 7.56s) on average to solve each problem.

At the end of the study, we selected one trial at random for compensation – a common procedure used to avoid the effects of reward accumulation during the task (Charness et al., 2016). If the participant had decided not to work on that trial, no bonus reward was received. If the participant decided to work for a reward on that trial, they would receive the reward offered on that trial. The decision of whether to work did not influence the rewards offered on future trials or pay-out of other members of the group. This information was emphasized in the instructions, and participants had to pass a comprehension check to ensure that they understood the details of the task.

The difference between the two onsite experiments was that in one experiment the participants knew the exchange rate between reward points offered and Great British Pounds (1 point was worth £0.04), in the other experiment it was unknown and said to differ on each trial (ranging from £0.001 to £0.08). The total bonus reward (after exchanging earned points from a selected trial to £) could range from £0 to £18.64 in Experiment 2 and from £0 to £37.28 in Experiment 3. We hypothesized that when the value of points was unknown, participants would rely more heavily on social context when deciding to work for a displayed reward. We replicated the core findings across both studies. Thus, we initially report results from the combined dataset, and then formally test if the effects differed in strength between both experiments. A separate analysis of each dataset is presented in the Supporting Information.

**Data analysis**

Although participants on average accepted 54% of reward offers in experiment 2 and 3, we found a considerable variability between participants, with some participants accepting/rejecting as little as just one offer, limiting inferences that can be drawn from a single participant. To account for this issue, as well as within-subject correlations of responses related to repeated measures in our design, we used Generalized Linear Mixed Effects (GLME) model approach, in which fixed effects describe the effect common for all

participants and random effects describe idiosyncrasies specific for an individual. The GLME model included decisions to work as the categorical dependent variable and assumed a binomial distribution of the dependent variable. The independent variables were unfairness (measured as Gini coefficient), rank (normalized to range from 0 to 1, for lowest and highest rank respectively), and reward magnitude (expressed as a power function, see below). All variables were standardized prior to the analysis. Following methodological recommendations by Barr and colleagues (Barr et al., 2013), all models included fixed and random effects for intercept and all independent variables.

Rank and unfairness were calculated as in Experiment 1. To account for a possibility of diminishing marginal utility of each additional awarded point, we tested if the effect of reward magnitude was better expressed as a linear or a power function (as it is in the prospect theory(Tversky & Kahneman, 1992)):

$$Reward\ magnitude\ utility = x_i^{\rho}$$

Where $x$ is the reward offer, and $\rho$ represents parameter describing the curvature of the reward function, ranging from 0 to 1 (at which point it is linear). To fit the above function, we estimated non-linear mixed-effects model with stochastic Expectation-Maximization algorithm (Delyon et al., 1999). The $\rho$ value maximizing the $R^2$ of the model describing the relationship between reward magnitude and motivation to work (including the variables listed in the section below) was equal to 0.43, suggesting a non-linear relationship between absolute reward and its value, and was subsequently used in all analyses.

We additionally tested if skewness of the distribution could separately influence participants' decisions, by including in the above model an Adjusted Pearson's Coefficient of Skewness, calculated as follows:

$$Adjusted\ Pearson's\ Coefficient\ of\ Skewness = \frac{\sqrt{n(n-1)}\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{(n-2)\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)^3}$$

Where $\bar{x}$ is the average reward offer, $n$ is the number of participants in the group, $x_i$ is the reward offer received by each person.

To illustrate the size of the effect of unfairness and rank we plotted predicted values of the above GLME model across different levels of unfairness (Fig 4A) and separately, across different ranks (Fig 4B), with the effect of trial number, rank (only for Fig 4A) and unfairness (only for Fig 4B) set to 0. To illustrate the effect of unfairness and rank in isolation from reward magnitude (Fig 4C), we estimated the probability of pursuing rewards on each trial from a GLME model including absolute reward and trial number (with other factors fixed to 0). We then calculated the residuals, by subtracting observed decisions and their predicted probability. We categorized residuals into 5 ranks and two levels of unfairness (based on the middle value of the tested range) and calculated the average residual value for each participant within each category and plotted the averages over participants within each category.

The fit of the model including rank and unfairness was compared to two other popular models describing the effects of inequality on the evaluation of reward: the adaptation model (Helson, 1964), and inequality aversion model (Fehr & Schmidt, 1999). All compared models included absolute value as one of the independent variables. Adaptation model additionally included the difference between absolute value of the offered reward and the average reward offered to all people in the group on a specific trial. Inequality aversion additionally model included advantageous and disadvantageous inequality, calculated as follows (Fehr & Schmidt, 1999):

$$Advantaegous\ inequality = \sum_{j=1}^{n} max\left|x_i - x_j, 0\right|$$

$$Disadvantageous\ inequality = \sum_{j=1}^{n} max\left|x_j - x_i, 0\right|$$

Where $x_i$ is an individual's payment offer and $x_j$ are payment offers received by other group members. All models were compared based on their Bayesian Information Criterion (BIC) which simultaneously assesses the model's fit, while penalizing it for its complexity.

To investigate if person's current emotional state mediated the effect of rank and unfairness on decisions to work, we used a multi-level mediation analysis approach (Kenny et al., 2003), which nests trial-level observations within upper-level units (individual participants), similarly to the GLME approach described above. The analysis was performed using M3 Mediation Toolbox for MATLAB (Wager et al., 2008). Bootstrapping approach, a non-parametric method based on resampling with replacement, was used to estimate the significance of the effects, using the standard 1000 samples (Hayes, 2009). To control for the fact that independent variables in our design were correlated and ensure that the conclusion of the mediation analysis relates specifically to the investigated variable, each mediation model was performed on residuals from a GLME model regressing out the effect of the variable not tested. That is regressing out trial number, and: (i) reward magnitude and rank for the mediation model describing the effect of unfairness, or (ii) reward magnitude and unfairness for the mediation model describing the effect of rank; on both feelings and decisions to work. Prior to the analysis, feelings ratings were transformed to range from 0 to 1, with 0 indicating a low score (i.e., very unhappy).

## RESULTS EXPERIMENTS 2 AND 3

### Opportunity gaps reduce the motivation to work.

Across two onsite experiments, participants chose to work on 54% of trials. To test whether the hypothesized factors influenced participants' choices, we used a generalized linear mixed-effects model (GLME) predicting decisions to work for reward on every trial from unfairness level of all offers, rank of individual's offered reward (from 1 to 5), and the absolute value of the offered reward (expressed as a power function to account for diminishing marginal utility; see Methods for details). Additionally, we examined if participants reacted to reward offers differently when the minority of individuals are at the top of the distribution and the majority at the bottom or vice versa, by including in the model the signed skewness of the distribution (measured by Adjusted Pearson's Coefficient of Skewness). The possible effect of fatigue was accounted for by including trial number. All three hypothesized factors significantly influenced decisions to work in exchange for rewards. In particular, the likelihood of pursuing rewards was greater when (i) unfairness was low ($\beta$ = -0.29, $p < 0.001$), (ii) rank was high ($\beta$ = 0.92, $p < 0.001$) and (iii) absolute reward was high ($\beta$ = 2.82, $p < 0.001$). In addition, the likelihood of pursuing rewards decreased over time ($\beta$ = -1.04, $p < 0.001$), presumably due to fatigue. Skewness of the distribution did not have a significant effect ($\beta$ = 0.01, $p = 0.92$).

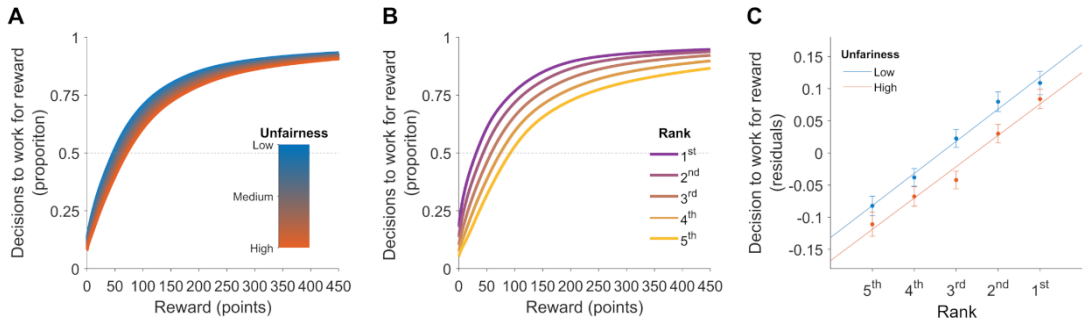


**Fig 4. Motivation to work is higher when the distribution of rewards is fair; the rank is high, and the absolute reward is high.** To illustrate the effect of factors influencing the motivation to work in repeated measures experiments, we plotted the probability of

participants' decision to work from a GLME model predicting choice from reward magnitude and either different levels of (A) unfairness or (B) rank. (C) We also plotted average residuals for the five rank categories and two levels of unfairness from a GLME model predicting choice just from absolute reward and trial number. We observe that participants are more likely to decide to work when (A, C) rewards are fairly distributed and (B, C) when the rank is high than low. Error bars = SEM.

To illustrate the impact of unfairness, we calculated each participant's probability of pursuing rewards at different levels of unfairness and reward magnitudes (based on the estimated fixed and random effects from a GLME model predicting decision to work only from these two factors, setting the other factors to 0). The estimated probabilities were then averaged over participants (Fig 4A). As can be observed, for the same reward magnitude, participants were more likely to work when unfairness was low rather than high. The indifference point (i.e., the reward magnitude for which participants choose to work with 50% probability) was 27.5 points greater for the highest level of unfairness than for the lowest level.

Next, we plotted the likelihood of pursuing rewards for each reward magnitude across the five offer ranks, using the same method as above. As can be observed in Fig 4B the likelihood of pursuing rewards was greater when the rank of the offer is high than when it was low for the same absolute value of the reward. For the lowest rank, participants required an additional 66.4 points to be indifferent on whether to pursue reward than for the highest rank.

To illustrate the effect of unfairness and rank in isolation from the reward magnitude, we plotted the residuals from the above GLME model with the effect of unfairness and rank set to 0. These residuals were then divided into five ranks and two levels of unfairness (high and low based on a median split; Fig 4C). This exercise demonstrates that participants were less likely to work when unfairness was high (red line) than low (blue line) across different ranks. Moreover, participants were more likely to work when the rank of their

reward offer was high than when it was low, across different levels of unfairness.

While large unfairness in the group had a negative effect on motivation, it may be that when looking downwards at the less fortunate, large unfairness might increase motivation. To test for this possibility, we added to the above GLME model two covariates for each subject and trial: the sum of distances between the participant and everyone below them (advantageous inequality) and the sum of the distances between the participant and everyone above them (disadvantageous inequality). While all three main effects from the original model remained significant (unfairness: $\beta = -0.33$, $p < 0.001$; rank: $\beta = 0.87$, $p < 0.001$; absolute reward: $\beta = 2.67$, $p < 0.001$), neither upward ($\beta = -0.04$, $p = 0.67$) nor downward ($\beta = -0.29$, $p = 0.08$) comparisons significantly influenced the willingness to work. In other words, while the relative ranking of a participant's pay offer affects motivation, as does the general level of unfairness, once we account for these two factors, having people's pay be at a greater distance from others' in either direction does not additionally impact their willingness to work.

Finally, we compared the original model to two well-known models in the literature that respectively describe the effect of relative value and inequality on utility: (i) the adaptation model, which is based on the assumption that people compare their income to an average value for their reference group (Helson, 1964), and (ii) the Fehr-Schmidt inequality aversion model, which assumes that people have a separate reaction to advantageous and disadvantageous inequality (Fehr & Schmidt, 1999). In both, we include absolute reward and trial number as covariates. Our original model (the 'rank-unfairness model') (BIC = 3661.9) outperformed both the adaptation model (BIC = 3765.8) and the Fehr-Schmidt inequality model (BIC = 3780.1), as well as models consisting of only rank (BIC = 3681.3), only unfairness (BIC = 3687.3), or only absolute reward (BIC = 3833.4). Together, the results suggest that high unfairness, low rank and low absolute reward all have significant, negative and

independent effects on the willingness to work and that both unfairness and relative value components are necessary to explain the reactions to unequal opportunities.

Across two experiments, we manipulated the level of uncertainty about the monetary value of points by either disclosing or not disclosing the exchange rate (£ per point). To test if the effects differed in these two cases, we added to our GLME model interaction effects between the version of the experiment and the three main factors: rank, unfairness, and absolute reward. We found that the effect of rank and unfairness was stronger when the value of points was unknown than when it was known (interaction with rank: $\beta = 0.97$, $p < 0.01$; interaction with unfairness: $\beta = -0.27$, $p = 0.019$), while remaining significant in both experiments (see Supporting Information). The effect of absolute reward was weaker when the value of points was unknown than when it was known (interaction between experiment version and absolute reward: $\beta = -1.7$, $p < 0.001$). This suggests that participants relied more heavily on social context when they were uncertain about monetary value.

**Feelings partially mediate the effects of opportunity gaps on decisions to exert effort.**

To examine whether feelings mediated the effects of opportunity gaps on decisions to work, we performed two multi-level mediation analyses. Each of the mediation analysis examined whether feelings mediate the effect of one of the factors identified above (i.e., rank or unfairness) while controlling for the absolute reward magnitude, trial number and the other factor.

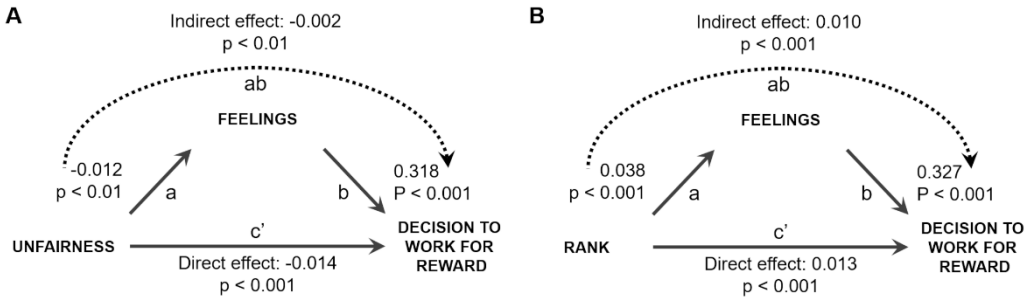


**Fig 5. Feelings partially mediate the effect of opportunity gaps on decisions to work for the reward**. We examined whether the effect of the two components of the motivational response to opportunity gaps, that is (A) unfairness and (B) rank, were mediated by feelings. In both cases, we controlled for the absolute reward, trial number and either rank (A) or unfairness (B) respectively. In both cases, we found a significant indirect effect and direct effect (which represents the influence of the given factor on decision to work, while controlling for the indirect effect), suggesting that feelings partially mediate the influence of each of the factors on decisions to work.

We found that the effects of unfairness and rank on decision to work were both partially mediated by feelings (see Fig 5). First, as we already reported, low unfairness and high rank were related to greater likelihood to work (total effect: unfairness: $\beta = -0.019$, $p < 0.001$; rank: $\beta = 0.029$, $p < 0.001$). This effect was partially mediated by feelings (path ab: unfairness: $\beta = -0.002$, $p < 0.001$; rank: $\beta = 0.010$, $p < 0.001$) with positive feelings related to low unfairness and high rank (path a: unfairness: $\beta = -0.012$, $p < 0.001$; rank: $\beta = 0.038$, $p < 0.001$). Additionally, feelings predicted decisions to work even when unfairness and rank were accounted for (path b: unfairness: $\beta = 0.318$, $p < 0.001$; rank: $\beta = 0.327$, $p < 0.001$). This suggests that incidental fluctuations of feelings, unrelated to task variables, also had a unique effect on the decision to work. Conversely, the two task related variables had direct effect on the decision to work that could not be accounted for by changes in feelings (path c': unfairness: $\beta = -0.014$, $p < 0.01$; rank: $\beta = 0.013$, $p < 0.001$).

## DISCUSSION

Circumstances beyond a person's control, such as socio-economic status at birth, often determine the rewards available to a person for their efforts. In the current study, we investigated how decisions to work are altered by a person's awareness that some people in their group were luckier than others in the rewards they were offered for performing the same task. We hypothesized that the motivation to work would be influenced by the violation of the fairness principle and relative valuation of rewards. Across three experiments, we found that unfair distribution of rewards between group members had a negative impact on the decision to work not only of disadvantaged individuals but also of advantaged individuals. Specifically, high unfairness was related to a reduction in the likelihood that participants agreed to work for their reward irrespective of the magnitude of their reward and their relative position in the distribution. This is despite such refusal reducing the likelihood of receiving a bonus while having no impact on the rewards received by others.

Second, the likelihood of agreeing to work in exchange for reward was reduced when the rank of the offer was low and vice versa (i.e., higher rank was related to greater motivation to work), irrespective of the actual magnitude of the offered reward. The third factor modulating motivation was the absolute reward itself. The fact that absolute reward magnitude exerted influence even when controlling for the level of unfairness and offer rank suggests that while people do care about the rewards of others, they only partially adapt to present social context when deciding whether to work (Burke et al., 2016).

We find that the rank-unfairness model outperformed the adaptation (Helson, 1964) and inequality-aversion models (Fehr & Schmidt, 1999) in explaining participants' reactions to opportunity-gaps. The adaptation model assumes that people focus on the difference between their reward and the average reward, while the inequality aversion model assumes that people

focus on two types of inequality: less heavily weighted advantageous and more heavily weighted disadvantageous inequality. The advantageous inequality is based on the absolute difference between a person's reward and all other worse rewards, and the disadvantageous inequality is based on the difference between a person's reward and all other better rewards. All three models predict an increase of motivation with increasing relative value, but the rank-unfairness model is based on ordinal rather than absolute comparisons. Predictions of these model substantially diverge for the effect of statistical dispersion: the rank-unfairness model predicts a uniform decrease of motivation with increased statistical dispersion across all different ranks, while the inequality aversion model predicts a greater drop of motivation for people with lower than higher ranks. On the other hand, the adaption model predicts that person at the top should always be more motivated by an increasing statistical dispersion, as statistical dispersion is associated with greater deviation of their reward from the mean. In both cases, the observed pattern of results is more consistent with the predictions of the rank-unfairness model than the alternatives.

By manipulating the unfairness of offers, offer's rank and absolute reward in the second and third experiment, we were able to dissociate the influence of each of the three factors within the same individual. By doing so, we overcome a difficulty in studying these variables in the "real-world", where individuals with different traits or experiences may populate different parts of the distribution (Gelissen & de Graaf, 2006) - making it difficult to isolate the influence of these components from factors correlating with them, such as negative effects of stereotypes on aspirations (Migheli, 2015; Riegle-Crumb et al., 2011) or differences in risk aversion (Guiso and Paiella, 2008). Together, these findings suggest that individuals who are offered less than others are disadvantaged not only because the absolute reward they can possibly obtain is lower, but also because they might suffer from a motivational cost that reduces the likelihood of pursuing the rewards that are within their reach. The

latter may be due to a lower relative value of their rewards and a demotivating effect of participating in a situation that seems unfair.

Importantly, because the decisions to work were made in private and did not affect others, the observed effect of unfairness on motivation cannot be attributed to reputation concerns (Engelmann & Fischbacher, 2009), reciprocity (Kube et al., 2012) or retribution motives (Suleiman, 1996). Instead, our results suggest that unfairness and rank exert their effect on motivation partially by influencing experienced feelings. We report a mediation that includes two links: the first is between each of the two factors (unfairness and low rank) and negative feelings; and the second between negative feelings and a reduction in the willingness to exert effort. As for the first link, high unfairness and low rank each triggered negative feelings even when controlling for the magnitude of the reward offered. The negative impact of opportunity gaps on feelings supports the notion that the perception of unfairness is reflected in emotional response (Rutledge, Berker, et al., 2016) and thus carries a cost to one's psychological well-being. The finding that rank influenced experienced feelings is consistent with studies showing that well-being measures are influenced by a person's standing relative to others (Boyce et al., 2010; Hagerty, 2000; Lyubomirsky & Ross, 1997).

The second link is between feelings and the willingness to work for the reward. Although the idea that unhappiness is related to low motivation is intuitive, there has not been conclusive evidence for it in healthy individuals ( for review see: Taris et al., 2014). Past studies have mostly examined the relationship between mood and performance level, rather than the decision to engage in effort altogether, and produced mixed results. While some researchers found a beneficial effect of positive mood induction on performance (Oswald et al., 2015), others found that positive and negative emotions can improve or impair performance depending on the nature of the task (Dreisbach, 2006; Dreisbach & Goschke, 2004; Gray, 2001; Phillips et al., 2002). With regards to the motivation to pursue rewards, we find that

unhappiness has a negative effect. Such an effect could be explained by the negative influence of bad mood on the perceived value of rewards, as suggested by previous experimental studies (Eldar & Niv, 2015; Huys et al., 2013). Alternatively, rather than playing a causal role, lower happiness in our study could simply index reaction to lower the subjective value of offered rewards (Rutledge et al., 2014). The effects of rank and unfairness were also observed in the first experiment, despite not asking participants about their current emotional state. This suggests that the influence of unfairness and rank on motivation is not conditional on prompted introspection.

The mediatory effect of feelings in a relationship between unfairness and willingness to work for reward was partial, suggesting that additional mechanisms drive the negative influence of unfairness on motivation. One such possibility is that participants use information about the social environment to resolve uncertainty about the value of their offers. In line with this suggestion, we found that in the condition in which the value of points was unknown, the effects of rank and unfairness were stronger than when the value of points was known.

Our study may have implications for people's decisions and behaviour outside the lab. We speculate that negative experiences caused by arbitrary reward disparities might contribute to higher prevalence be one reason why disadvantaged individuals are more likely to suffer from anxiety and depression (González et al., 2010; Lee et al., 2017; Piccinelli & Wilkinson, 2000). Furthermore, decreased motivation caused by unfairness and low relative position might make upward mobility particularly difficult, contributing to sustained poverty among disadvantaged groups (Elmelech & Lu, 2004; Findlay & Wright, 1996; Uhrig, 2015). As such, the motivational phenomenon described in this study constitutes another example of a poverty-trap, that is a situation where having worse prospects triggers additional mechanisms ensuring that a person remains poor. It also suggests that any observed signs of decreased motivation among disadvantaged groups might be situational,

rather than stemming from an individual's characteristics and could be a potential target of interventions.

The instructions of the studies made clear to participants that they had no control over the magnitude of the rewards offered. In contrast, in many everyday situations, there is ambiguity about the role of randomness in success. Previous studies have shown that in such ambiguous situations, those who are advantaged are more likely to assume that their economic position is a result of talent and effort, while those who are disadvantaged assume it is a result of external circumstances (Hunt, 2004; Kluegel & Smith, 1986). It remains to be tested whether similar effects to those reported here would be observed in such situations.

While past studies have suggested that people are generally averse to unfair distributions of rewards, here we uncover their consequences beyond distribution preferences (Dawes et al., 2007; Fehr & Schmidt, 1999) or impact on the affective state (Rutledge, Berker, et al., 2016; Tricomi et al., 2009). We show that unequal opportunities have a negative influence on the motivation to work for the reward of not only disadvantaged individuals but also of others around them. Our findings provide an empirical framework for considering the impact of opportunity gaps on individuals, organizations, and societies, suggesting they can trigger psychological dynamics that hurt the productivity of all involved.

# SUPPLEMENTARY INFORMATION

**S1 Table. Distributions of payment offers between participants (expressed in points) for each trial presented in Experiments 2 and 3.**

| Rank | | | | |
|---|---|---|---|---|
| 5th | 4th | 3rd | 2nd | 1st |
| 25 | 31 | 36 | 42 | 54 |
| 22 | 29 | 35 | 42 | 59 |
| 19 | 27 | 34 | 43 | 65 |
| 17 | 25 | 33 | 44 | 73 |
| 15 | 23 | 32 | 45 | 82 |
| 13 | 21 | 31 | 47 | 94 |
| 10 | 19 | 30 | 48 | 110 |
| 8 | 17 | 29 | 50 | 129 |
| 7 | 15 | 28 | 52 | 155 |
| 5 | 14 | 27 | 55 | 189 |
| 34 | 44 | 51 | 60 | 79 |
| 30 | 41 | 50 | 62 | 88 |
| 27 | 39 | 49 | 64 | 98 |
| 23 | 36 | 48 | 66 | 111 |
| 20 | 34 | 47 | 68 | 127 |
| 17 | 31 | 46 | 71 | 146 |
| 15 | 29 | 46 | 74 | 170 |
| 12 | 26 | 45 | 77 | 201 |
| 10 | 24 | 44 | 81 | 242 |
| 7 | 21 | 43 | 87 | 297 |
| 48 | 64 | 76 | 90 | 119 |
| 43 | 60 | 75 | 93 | 133 |
| 38 | 57 | 74 | 96 | 150 |
| 33 | 53 | 73 | 100 | 170 |
| 29 | 50 | 72 | 104 | 196 |
| 25 | 47 | 71 | 109 | 226 |
| 21 | 43 | 70 | 114 | 265 |
| 18 | 40 | 69 | 120 | 315 |
| 15 | 37 | 68 | 127 | 379 |
| 12 | 33 | 67 | 136 | 466 |
| 25 | 37 | 43 | 48 | 54 |
| 22 | 39 | 46 | 52 | 59 |
| 19 | 41 | 50 | 57 | 65 |
| 17 | 46 | 57 | 65 | 73 |
| 15 | 52 | 65 | 74 | 82 |
| 13 | 60 | 76 | 86 | 94 |
| 10 | 72 | 90 | 101 | 110 |
| 8 | 87 | 108 | 120 | 129 |
| 7 | 110 | 134 | 147 | 155 |
| 5 | 139 | 167 | 180 | 189 |
| 34 | 53 | 62 | 69 | 79 |
| 30 | 56 | 68 | 77 | 88 |
| 27 | 61 | 76 | 86 | 98 |
| 23 | 68 | 86 | 98 | 111 |
| 20 | 79 | 100 | 113 | 127 |
| 17 | 92 | 117 | 132 | 146 |
| 15 | 111 | 139 | 156 | 170 |

| | | | | |
|---|---|---|---|---|
| 12 | 136 | 168 | 187 | 201 |
| 10 | 171 | 208 | 228 | 242 |
| 7 | 217 | 261 | 283 | 297 |
| 48 | 77 | 91 | 103 | 119 |
| 43 | 83 | 101 | 116 | 133 |
| 38 | 92 | 114 | 131 | 150 |
| 33 | 103 | 130 | 150 | 170 |
| 29 | 121 | 153 | 175 | 196 |
| 25 | 142 | 180 | 204 | 226 |
| 21 | 172 | 216 | 243 | 265 |
| 18 | 213 | 264 | 293 | 315 |
| 15 | 267 | 326 | 357 | 379 |
| 12 | 342 | 411 | 445 | 466 |

**S2 Table. Distributions of payment offers between participants (expressed in pence) for each condition in Experiment 1.**

| Offer rank | Fair distribution | | | | | Unfair distribution | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 1st | 8 | 11 | 14 | 17 | **24** | 1 | 3 | 5 | 8 | **24** |
| 2nd | 12 | 16 | 19 | **24** | 34 | 4 | 8 | 13 | **24** | 73 |
| 3rd | 16 | 21 | **24** | 32 | 45 | 8 | 15 | **24** | 45 | 136 |
| 4th | 18 | **24** | 29 | 36 | 51 | 12 | **24** | 41 | 74 | 225 |
| 5th | **24** | 32 | 39 | 48 | 69 | **24** | 46 | 80 | 144 | 435 |

**S3 Table. Influence of unfairness, rank and absolute reward on experienced feelings.** GLME model predicting self-reported feelings.

| | Coefficient (SE) | T-stat | P-value |
|---|---|---|---|
| **Intercept** | 0.45(0.015) | 29.94 | < 0.0001 |
| **Trial** | -0.052(0.006) | -8.44 | < 0.0001 |
| **Absolute reward** | 0.095(0.008) | 11.85 | < 0.0001 |
| **Rank** | 0.070(0.009) | 8.08 | < 0.0001 |
| **Unfairness** | -0.011(0.004) | -3.60 | < 0.0001 |

**S4 Table**. **GLME model predicting decisions to pursue rewards in Experiment 2 (value of points known).**

| | Coefficient (SE) | T-stat | P-value |
|---|---|---|---|
| **Intercept** | 1.10(0.60) | 1.83 | < 0.01 |
| **Trial** | -1.45(0.17) | -8.37 | < 0.0001 |
| **Absolute reward** | 4.33(0.42) | 10.27 | < 0.0001 |
| **Rank** | 0.37(0.12) | 3.13 | < 0.0001 |
| **Unfairness** | -0.14(0.09) | -1.98 | 0.058 |

**S5 Table. GLME model of decisions to pursue rewards, Experiment 3 (value of points unknown).**

| | Coefficient (SE) | T-stat | P-value |
|---|---|---|---|
| **Intercept** | 0.45 (0.35) | 1.29 | 0.20 |
| **Trial** | -0.73(0.13) | -9.77 | < 0.0001 |
| **Absolute reward** | 1.70(0.22) | 7.96 | < 0.0001 |
| **Rank** | 1.33(0.14) | 9.40 | < 0.0001 |
| **Unfairness** | -0.41 (0.09) | -4.40 | < 0.0001 |

## REFERENCES


Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bault, N., Joffily, M., Rustichini, A., & Coricelli, G. (2011). Medial prefrontal cortex and striatum mediate the influence of social comparison on the decision process. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(38), 16044–16049. JSTOR.

Blau, F. D., & Kahn, L. M. (2007). The Gender Pay Gap: Have Women Gone as Far as They Can? *Academy of Management Perspectives*, *21*(1), 7–23.

Boliver, V. (2013). How fair is access to more prestigious UK universities? *The British Journal of Sociology*, *64*(2), 344–364. https://doi.org/10.1111/1468-4446.12021

Boyce, C. J., Brown, G. D. A., & Moore, S. C. (2010). Money and Happiness: Rank of Income, Not Income, Affects Life Satisfaction. *Psychological Science*, *21*(4), 471–475. JSTOR.

Burke, C. J., Baddeley, M., Tobler, P. N., & Schultz, W. (2016). Partial Adaptation of Obtained and Observed Value Signals Preserves Information about Gains and Losses. *The Journal of Neuroscience*, *36*(39), 10016–10025. https://doi.org/10.1523/JNEUROSCI.0487-16.2016

Charness, G., Gneezy, U., & Halladay, B. (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization*, *131*, 141–150. https://doi.org/10.1016/j.jebo.2016.08.010

Chetty, R., & Hendren, N. (2018). The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects. *The Quarterly Journal of Economics*, *133*(3), 1107–1162. https://doi.org/10.1093/qje/qjy007

Clark, A. E., & D'Ambrosio, C. (2015). Chapter 13 - Attitudes to Income Inequality: Experimental and Survey Evidence. In A. B. Atkinson & F. Bourguignon (Eds.), *Handbook of Income Distribution* (Vol. 2, pp. 1147–1208). Elsevier. https://doi.org/10.1016/B978-0-444-59428-0.00014-X

Crawford, C., Macmillan, L., & Vignoles, A. (2014). *Progress made by high-attaining children from disadvantaged backgrounds: Research report*. Social Mobility and Child Poverty Commission. http://dera.ioe.ac.uk/20433/1/High_attainers_progress_report_final.pdf

Dawes, C. T., Fowler, J. H., Johnson, T., McElreath, R., & Smirnov, O. (2007). Egalitarian motives in humans. *Nature*, *446*(7137), 794–796. https://doi.org/10.1038/nature05651

Delyon, B., Lavielle, M., & Moulines, E. (1999). Convergence of a Stochastic Approximation Version of the EM Algorithm. *The Annals of Statistics*, *27*(1), 94–128. JSTOR.

Dreisbach, G. (2006). How positive affect modulates cognitive control: The costs and benefits of reduced maintenance capability. *Brain and Cognition*, *60*(1), 11–19. https://doi.org/10.1016/j.bandc.2005.08.003

Dreisbach, G., & Goschke, T. (2004). How positive affect modulates cognitive control: Reduced perseveration at the cost of increased distractibility. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*(2), 343–353. https://doi.org/10.1037/0278-7393.30.2.343

Duncan, G. J., & Magnuson, K. (2012). Socioeconomic status and cognitive functioning: Moving from correlation to causation. *Wiley Interdisciplinary Reviews. Cognitive Science*, *3*(3), 377–386. https://doi.org/10.1002/wcs.1176

Eldar, E., & Niv, Y. (2015). Interaction between emotional state and learning underlies mood instability. *Nature Communications*, *6*(1), 1–10. https://doi.org/10.1038/ncomms7149

Elmelech, Y., & Lu, H.-H. (2004). Race, ethnicity, and the gender poverty gap. *Social Science Research*, *33*(1), 158–182. https://doi.org/10.1016/S0049-089X(03)00044-9

Engelmann, D., & Fischbacher, U. (2009). Indirect reciprocity and strategic reputation building in an experimental helping game. *Games and Economic Behavior*, *67*(2), 399–407. https://doi.org/10.1016/j.geb.2008.12.006

Fehr, E., & Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. https://doi.org/10.1162/003355399556151

Findlay, J., & Wright, R. E. (1996). Gender, Poverty and the Intra-Household Distribution of Resources. *Review of Income and Wealth*, *42*(3), 335–351. https://doi.org/10.1111/j.1475-4991.1996.tb00186.x

Gelissen, J., & de Graaf, P. M. (2006). Personality, social background, and occupational career success. *Social Science Research*, *35*(3), 702–726. https://doi.org/10.1016/j.ssresearch.2005.06.005

González, H. M., Tarraf, W., Whitfield, K. E., & Vega, W. A. (2010). The epidemiology of major depression and ethnicity in the United States. *Journal of Psychiatric Research*, *44*(15), 1043–1051. https://doi.org/10.1016/j.jpsychires.2010.03.017

Gray, J. R. (2001). Emotional modulation of cognitive control: Approach-withdrawal states double-dissociate spatial from verbal two-back task performance. *Journal of Experimental Psychology. General*, *130*(3), 436–452.

Guiso, L., & Paiella, M. (2008). Risk Aversion, Wealth, and Background Risk. *Journal of the European Economic Association*, *6*(6), 1109–1150. JSTOR.

Hagerty, M. R. (2000). Social comparisons of income in one's community: Evidence from national surveys of income and happiness. *Journal of Personality and Social Psychology*, *78*(4), 764–771.

Hayes, A. F. (2009). Beyond Baron and Kenny: Statistical Mediation Analysis in the New Millennium. *Communication Monographs*, *76*(4), 408–420. https://doi.org/10.1080/03637750903310360

Helson, H. (1964). *Adaptation-level theory* (Vol. xvii). Harper & Row.

Hunt, M. O. (2004). Race/Ethnicity and Beliefs about Wealth and Poverty. *Social Science Quarterly*, *85*(3), 827–853. JSTOR.

Huys, Q. J., Pizzagalli, D. A., Bogdan, R., & Dayan, P. (2013). Mapping anhedonia onto reinforcement learning: A behavioural meta-analysis. *Biology of Mood & Anxiety Disorders*, *3*(1), 1–16. https://doi.org/10.1186/2045-5380-3-12

Kenny, D. A., Korchmaros, J. D., & Bolger, N. (2003). Lower level mediation in multilevel models. *Psychological Methods*, *8*(2), 115–128.

Kluegel, J. R., & Smith, E. R. (1986). *Beliefs About Inequality: Americans' Views of What Is and What Ought to Be*. Transaction Publishers.

Kube, S., Maréchal, M. A., & Puppe, C. (2012). The Currency of Reciprocity: Gift Exchange in the Workplace. *American Economic Review*, *102*(4), 1644–1662. https://doi.org/10.1257/aer.102.4.1644

Lazear, E. P., & Shaw, K. L. (2009). *The Structure of Wages: An International Comparison*. University of Chicago Press.

Lee, C., Oliffe, J. L., Kelly, M. T., & Ferlatte, O. (2017). Depression and Suicidality in Gay Men: Implications for Health Care Providers.

*American Journal of Men's Health*, *11*(4), 910–919. https://doi.org/10.1177/1557988316685492

Lyubomirsky, S., & Ross, L. (1997). Hedonic consequences of social comparison: A contrast of happy and unhappy people. *Journal of Personality and Social Psychology*, *73*(6), 1141–1157.

McCall, L., Burk, D., Laperrière, M., & Richeson, J. A. (2017). Exposure to rising inequality shapes Americans' opportunity beliefs and policy support. *Proceedings of the National Academy of Sciences*, *114*(36), 9593–9598. https://doi.org/10.1073/pnas.1706253114

Migheli, M. (2015). Gender at work: Incentives and self-sorting. *Journal of Behavioral and Experimental Economics*, *55*, 10–18. https://doi.org/10.1016/j.socec.2014.12.005

Milanovic, B. (2014). Global Inequality of Opportunity: How Much of Our Income Is Determined by Where We Live? *The Review of Economics and Statistics*, *97*(2), 452–460. https://doi.org/10.1162/REST_a_00432

Oishi, S., Kesebir, S., & Diener, E. (2011). Income Inequality and Happiness. *Psychological Science*, *22*(9), 1095–1100. https://doi.org/10.1177/0956797611417262

Oswald, A. J., Proto, E., & Sgroi, D. (2015). Happiness and Productivity. *Journal of Labor Economics*, *33*(4), 789–822. https://doi.org/10.1086/681096

Phillips, L. H., Bull, R., Adams, E., & Fraser, L. (2002). Positive mood and executive function: Evidence from stroop and fluency tasks. *Emotion (Washington, D.C.)*, *2*(1), 12–22.

Piccinelli, M., & Wilkinson, G. (2000). Gender differences in depression: Critical review. *The British Journal of Psychiatry*, *177*(6), 486–492. https://doi.org/10.1192/bjp.177.6.486

Pinkovskiy, M., & Sala-i-Martin, X. (2009). *Parametric Estimations of the World Distribution of Income* (Working Paper No. 15433; Working Paper Series). National Bureau of Economic Research. https://doi.org/10.3386/w15433

Pluchino, A., Biondo, A. E., & Rapisarda, A. (2018). Talent versus luck: The role of randomness in success and failure. *Advances in Complex Systems*, *21*(03n04), 1850014. https://doi.org/10.1142/S0219525918500145

Riegle-Crumb, C., Moore, C., & Ramos-Wada, A. (2011). Who wants to have a career in science or math? Exploring adolescents' future aspirations by gender and race/ethnicity. *Science Education*, *95*(3), 458–476. https://doi.org/10.1002/sce.20431

Rutledge, R. B., Berker, A. O. de, Espenhahn, S., Dayan, P., & Dolan, R. J. (2016). The social contingency of momentary subjective well-being. *Nature Communications*, *7*(1), 1–8. https://doi.org/10.1038/ncomms11825

Rutledge, R. B., de Berker, A. O., Espenhahn, S., Dayan, P., & Dolan, R. J. (2016). The social contingency of momentary subjective well-being. *Nature Communications*, *7*(1), 1–8. https://doi.org/10.1038/ncomms11825

Rutledge, R. B., Skandali, N., Dayan, P., & Dolan, R. J. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, *111*(33), 12252–12257. https://doi.org/10.1073/pnas.1407535111

Silberzahn, R., & Uhlmann, E. L. (2013). It Pays to Be Herr Kaiser: Germans With Noble-Sounding Surnames More Often Work as Managers Than as Employees. *Psychological Science*, *24*(12), 2437–2444. https://doi.org/10.1177/0956797613494851

Starmans, C., Sheskin, M., & Bloom, P. (2017). Why people prefer unequal societies. *Nature Human Behaviour*, *1*(4), 1–7. https://doi.org/10.1038/s41562-017-0082

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. *Cognitive Psychology*, *53*(1), 1–26. https://doi.org/10.1016/j.cogpsych.2005.10.003

Suleiman, R. (1996). Expectations and fairness in a modified Ultimatum game. *Journal of Economic Psychology*, *17*(5), 531–554. https://doi.org/10.1016/S0167-4870(96)00029-3

Taris, T. W., Schaufeli, W. B., & Schaufeli, W. B. (2014, November 13). *Individual well-being and performance at work: A conceptual and theoretical overview*. Well-Being and Performance at Work; Psychology Press. https://doi.org/10.4324/9781315743325-6

Thiele, T., Pope, D., Singleton, A., Snape, D., & Stanistreet, D. (2017). Experience of disadvantage: The influence of identity on engagement

in working class students' educational trajectories to an elite university. *British Educational Research Journal*, *43*(1), 49–67. https://doi.org/10.1002/berj.3251

Tricomi, E., Balleine, B. W., & O'Doherty, J. P. (2009). A specific role for posterior dorsolateral striatum in human habit learning. *The European Journal of Neuroscience*, *29*(11), 2225–2232. https://doi.org/10.1111/j.1460-9568.2009.06796.x

Tversky, A., & Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. https://doi.org/10.1007/BF00122574

Uhrig, S. N. (2015). SEXUAL ORIENTATION AND POVERTY IN THE UK: A REVIEW AND TOP-LINE FINDINGS FROM THE UK HOUSEHOLD LONGITUDINAL STUDY. *Journal of Research in Gender Studies*, *5*(1), 23–72.

Wager, T. D., Davidson, M. L., Hughes, B. L., Lindquist, M. A., & Ochsner, K. N. (2008). Prefrontal-subcortical pathways mediating successful emotion regulation. *Neuron*, *59*(6), 1037–1050. https://doi.org/10.1016/j.neuron.2008.09.006

# Chapter 3

# Subjective Perceptions Of Inequality And Their (Dis)agreement With Normative Axioms Underlying Inequality Measurement.

**Filip Gesiarz[*1], Jan-Emmanuel De Neve[2], Tali Sharot[*1]**

[1]Affective Brain Lab, Department of Experimental Psychology, University College London, London, UK

[2]Saïd Business School, University of Oxford, Oxford, UK

## ABSTRACT

How do we form judgments about the levels of inequality around us? Despite a recent surge of interest in the discrepancies between actual and perceived inequality in one's country, we still lack the basic understanding of how people make inequality judgments when presented with actual information about incomes. Here we study such judgments and document how they violate several normative principles underlying inequality measures used in economics. In an experiment, we expose participants to 60 different income distributions and ask them to evaluate their inequality. We demonstrate that people violate the anonymity principle, by being affected by their position in the distribution, the scale-independence principle, by being affected by the size of the economy, and the additivity principle, by being insensitive to the addition of incomes that transforms positively skewed distributions into negatively skewed ones. We find partial support for the transfer principle, showing that people are more sensitive to transfers of money to the poor than to the rich. Out of all tested non-parametric measures of inequality, the mean absolute difference between incomes most closely approximated subjective perception. To integrate these findings, we develop a new index of subjective inequality that fits data better than any other commonly used measure. Our findings provide a quantitative characterization of principles governing subjective inequality, with potential implications to different lines of research, including numerical perception, contextual influences on valuation, and welfare economics.

## INTRODUCTION

How high is the inequality in one's own country? It has been shown that to answer this question people extrapolate from their local environments (Cruces, Perez-Truglia, and Tetaz, 2013). However, we still lack a basic understanding of how a person can form a judgment about inequality based on just a few examples of incomes that they know about, and what criteria do they use to make such evaluations. The aim of this study is to quantitively characterize these judgments and show how subjective perceptions deviate from objective measures of inequality. To measure the effect of inequality on society, we first have to decide how to quantify it, and the current study provides many insights that can stimulate the debate on what is important for the evaluations of inequality from the perspective of the public.

We build on the seminal work of Amiel and Cowell (1999), who demonstrated that inequality judgments involving pair-wise comparisons of income distributions often violate common assumptions underlying inequality measurement in economics. We extend these findings, using a methodology that relies on continuous rather than categorical judgments and identifying measures that most closely approximate the lay perceptions of inequality. We also develop the Subjective Equality Index, based on the Gini coefficient, that is consistent with principles governing how people perceive statistical dispersion, as opposed to measures based on normative assumptions.

Many measures of inequality in economics, such as the Gini coefficient, are based on normative axioms that were developed to measure statistical dispersion (Dalton, 1920; Blackorby, Bossert, Donaldson, 1999). As such, there are neutral with regards to fairness or social welfare considerations. It is possible that lay perceptions of inequality might deviate from objective measures of inequality precisely because lay perceptions might be influenced by self-regarding or other-regarding motives.

First of these axioms is the anonymity principle, according to which the identities of people receiving income should not affect the estimation of

inequality. In contrast, we hypothesize that personal position in the distribution might bias equality judgments, favoring equality perception whenever a person is advantaged and favoring inequality perception whenever a person is disadvantaged, violating the anonymity principle. Such bias could arise from many different mechanisms that incorporate seemingly irrelevant information into judgments, such as anchor effects (Markovsky, 1988), motivated reasoning (Kunda, 1990), or using affect induced by rank as information (van den Bos, 2003).

The second axiom is scale-independence, according to which a proportional increase of all incomes by some factor should not change the estimation of inequality. In measures of inequality used in economics, this assumption is implemented by dividing a measure of dispersion by some normalizing factor (thereafter referred to as 'normalization'). In the case of the Gini coefficient, for example, it is twice the mean of all incomes. Subjective inequality judgments, however, do not necessarily exhibit scale-independence (Amiel and Cowell, 1999). We test if scale-(in)dependence differs between individuals and if it falls on a continuum between full insensitivity and fully proportional scaling to the size of the economy, rather than being characterized by categorical extremes.

The last axiom is the transfer-principle, according to which any transfer of income from a richer to a poorer person should decrease inequality. It has been found that lay perceptions of inequality are consistent with this principle only in the case of transfers from the very rich, to the less rich, but not in the case of transfers from the poor to the poorer (Amiel and Cowell, 1999). This finding implies a differential sensitivity to changes in incomes in different parts of the distribution: subjective perception of inequality might be different depending on from whom the money is taken away from and to whom it is given to. We evaluate this possibility statistically by fitting an extended version of the Gini index (Yitzhaki, 1983) – an inequality measure that additionally

quantifies the weight put on each part of the distribution, allowing for differential sensitivity to changes in income across the distribution.

We also address the additivity principle, which is not part of the normative set of axioms underlying inequality measurement but is one of the assumptions in the social welfare theories (Temkin, 1983). According to this principle, increasing income of any person that is not the target of income aspirations (usually assumed to be the person at the top), should increase social welfare. However, it has been observed that gradually adding incomes to people in the middle of the distribution, starting from the right side of the distribution, resulted in a U-shape pattern of inequality judgments (Amiel and Cowell, 1999). Such a pattern is inconsistent with scale-independence and additivity principles, which would predict a steady decrease. Here we suggest that the above pattern arises due to the insensitivity of subjective perception to skewness. More specifically, we predict that distributions that are a mirror-image of each other in terms of skewness (while having the same range of incomes and standard deviation) will be considered similarly unequal. We also test if the violation of the additivity principle replicates in a situation where distributions are presented separately and in random order, avoiding direct contrasts between them – a confound noted by the authors of the previous study (Amiel and Cowell, 1999).

Our approach introduces several methodological advances. We test income distributions that participants had an opportunity to experience, rather than hypothetical situations. The distributions in our experiment are presented sequentially and in a random order, avoiding well-documented biases related to simultaneous presentation (Kahneman and Thaler, 2006; Read et al., 2001) or order effects (Mantonakis, 2009). The random and sequential presentation also avoids highlighting the features of the distributions that the researcher is interested in studying, minimizing the risk of participants conforming with inferred expectations. We use continuous rather than categorical subjective judgments, that allow for a more sensitive assessment of factors influencing

the perception of inequality. We supplement our findings with statistical analysis, extending so far mostly qualitative characterization of subjective perceptions of inequality to a more quantitative understanding of the phenomenon (Amiel and Cowell, 1999). Finally, we compare different inequality measures used in economics and identify which of them most closely match subjective perceptions of inequality. We also introduce the Subjective Equality Index - an inequality measure based on the modified Gini coefficient that exhibits many characteristics of the subjective perception of inequality and fits participants' data better than any other tested measure.

## METHODS

### Participants

We recruited one hundred and ten participants from the University College London subject pool to participate in an experiment (mean age 21.5[2.5], age range 18 – 35; 64% women). The experiment immediately proceeded experiments 2 and 3 from Chapter 2, and was presented as a second part of the study. For further details about the sample, please see 'Participants' section for study 2 and 3 in Chapter 2.

### Behavioral task

In the experiments described in Chapter 2, participants had an opportunity to experience different reward distributions. During this part, they were shown 5 reward offers representing 60 different reward distributions over 60 trials. Rewards were expressed in points and were exchanged into monetary rewards after the end of the experiment. On each trial, one reward out of five was assigned to a participant at random. The participant then had to decide if they accept the reward offer. If they accepted it, they had to perform a simple math task to earn their reward. If they rejected it, they proceeded to the next trial.

In the second part of the experiment (**Figure 1A**), participants were again presented with the same reward distributions as in the first part. This time, however, instead of deciding to accept or reject the reward offer, they were asked to judge how equal/unequal each distribution is, by clicking on a continuous sliding scale.

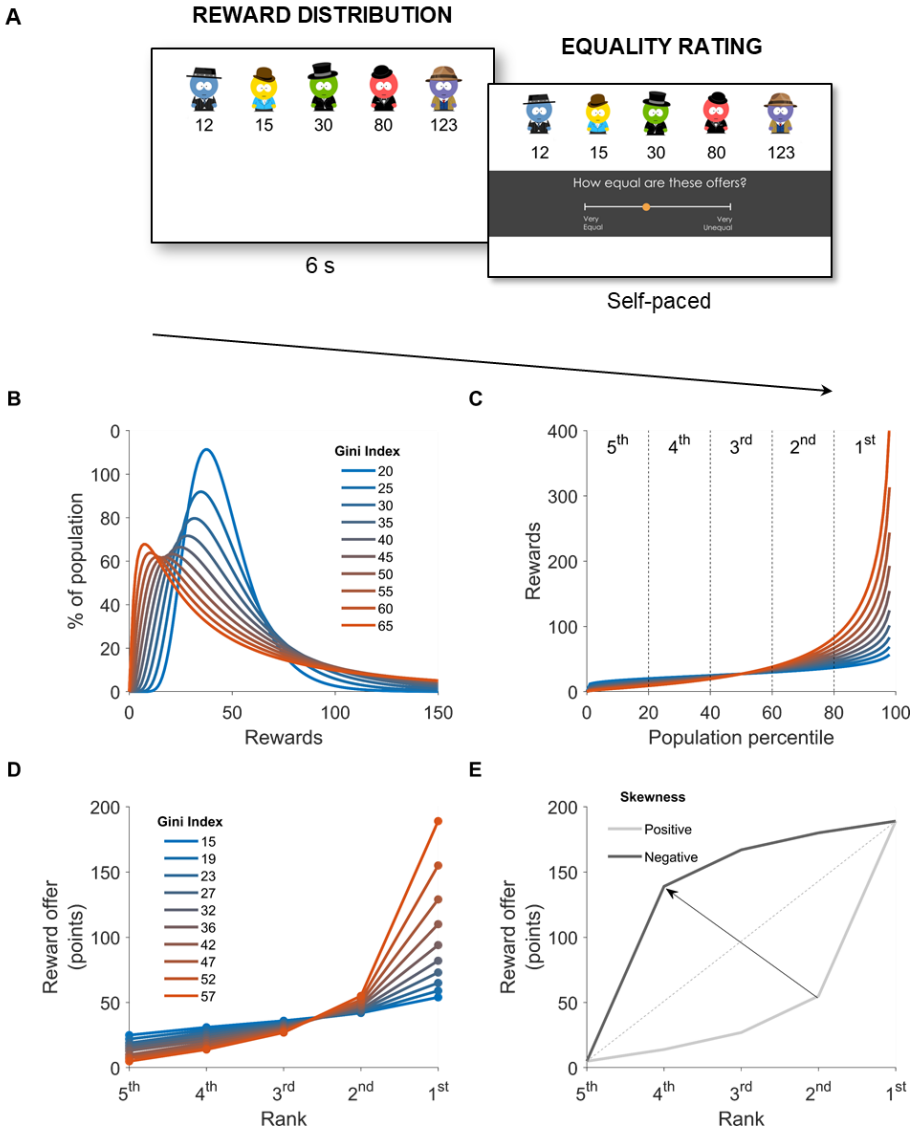


**Figure 1. Behavioral task and Reward distributions. A**) The task consisted of 60 trials during which participants, after viewing a reward distribution for 6 seconds, had to evaluate how equal are the presented rewards offers. These reward distributions were experienced by participants in the same form and order in a preceding task, in which participants had to decide if they reject or accept the presented offer that was assigned to them at random, and marked among 5 other offers by personal cartoon avatar. **B**) Out of 60 distributions, 30 income distributions were based on a log-normal probability density function (corresponding to 10

levels of Gini index uniformly distributed between 20 and 65, with 3 different median values). Log-normal distribution approximates reward distributions encountered in the real world, such as income distributions within countries (Pinkovskiy and Sala-i-Martin, 2009) and companies (Lazear and Shaw, 2007). For illustration purposes, figures A, B, and C show only ten of these income distributions based on only one scale value. **C**) To generate rewards representative for the above distributions, we used an inverse cumulative density function of these distributions, which assigns maximal income value earned by each percent of the population. **D**) We next took an average income from each quintile of this function, with the exclusion of the top 1 percentile, resulting in five representative values for each trial. The inequality of reward offers used in the analysis was quantified based on these five values. **E**) We transformed values from positively skewed distribution to create additional 30 negatively skewed reward distributions. The resulting distributions had the same range and standard deviation of rewards as the positively skewed distributions.

## Distribution of reward offers

We created 60 different distributions in total and presented them in random order. We generated 30 reward distributions based on a log-normal probability density function (**Figure 1B**). We chose log-normal distribution as it fits closely to real-world income structures within firms (Lazear and Shaw, 2016) and countries (Pinkovskiy and Sala-i-Martin, 2009). To vary the levels of reward magnitude range and statistical dispersion we used a combination of 3 different scale parameters (0.55, 1, 1.45) and ten different standard deviations, corresponding to values of the Gini coefficient varying uniformly from 20 to 65 (**Figure 2**), resulting in 30 different distributions. Log-normal distributions are always positively skewed. To generalize our findings, we also included 30 negatively skewed distributions that were a mirror-image of the positively skewed distributions by applying the following transformation of representative values (**Figure 1E**):

$$x_{positive} = \{x_1, x_2, x_3, x_4, x_5\}$$

$$x_{negative} = \left|x_{positive} - max\left(x_{positive}\right)\right| + \min\left(x_{positive}\right)$$

Where $x_n$ is subject $n$ payment offer in each trial, $x_{positive}$ and $x_{negative}$ are payment offers of all participants in trials with positively and negatively skewed distributions, respectively.

To generate reward offers representative of the above distributions, we used an inverse cumulative density function of these distributions (**Figure 1C**), which assigns maximal pay value earned by each percent of the population. We next took an average pay from the subsequent 20 percentiles of this function, with the exclusion of the top 1 percentile, resulting in 5 values reflecting an average pay of each 20% of the population (**Figure 1D**). We excluded the last percentile as it approaches infinity. Unfairness was quantified based on these five representative values. To introduce variability to the middle pay (that otherwise would be the same for all distributions generated from the same median value) we additionally subtracted a number between 0 and 9 from each representative value in each distribution (in each distribution the same number was subtracted for each value).

### Model fitting

For the purposes of model comparisons, the models of subjective perception of inequality were fit individually to each participant using the least-squares method and fmincon function in MATLAB. To avoid convergence of the fitting algorithm at the local minimum, each fitting procedure was re-run 100 times with random starting parameter values. The final parameter estimates were chosen out of all 100 fitting procedures based on the highest $R^2$ value.

## RESULTS

### Subjective perception of inequality violates the anonymity principle

We start our investigation with the principle of anonymity. According to this principle, the identity of the person receiving the income should not

influence the estimation of inequality (**Figure 2A**). We hypothesized that participants might be biased in their estimations of inequality by their position in the distribution. To test this hypothesis, we created a linear mixed-effects model predicting subjective inequality rating from a person's income rank in the distribution (centered at the middle rank) on each trial. The model also included a random intercept for each participant, to account for inter-trial correlations of equality ratings between trials.

We find that the higher the rank of the received offer, the more likely participants perceived distribution as equal ($\beta = 0.013$, $p < 0.001$). The plot of average equality ratings for each rank reveals that the effect was particularly strong for the lowest rank (**Figure 2A**, third panel). Indeed, including a dummy variable indicating if a person had the lowest rank on a particular trial suggests that the effect is primarily driven by the reaction to being at the bottom of the distribution (bottom rank dummy: $\beta = -0.026$, $p < 0.01$; rank: $\beta = 0.006$, $p = 0.11$). These results suggest that people are not impartial in their inequality judgments when they are personally affected by the situation, violating the anonymity principle.

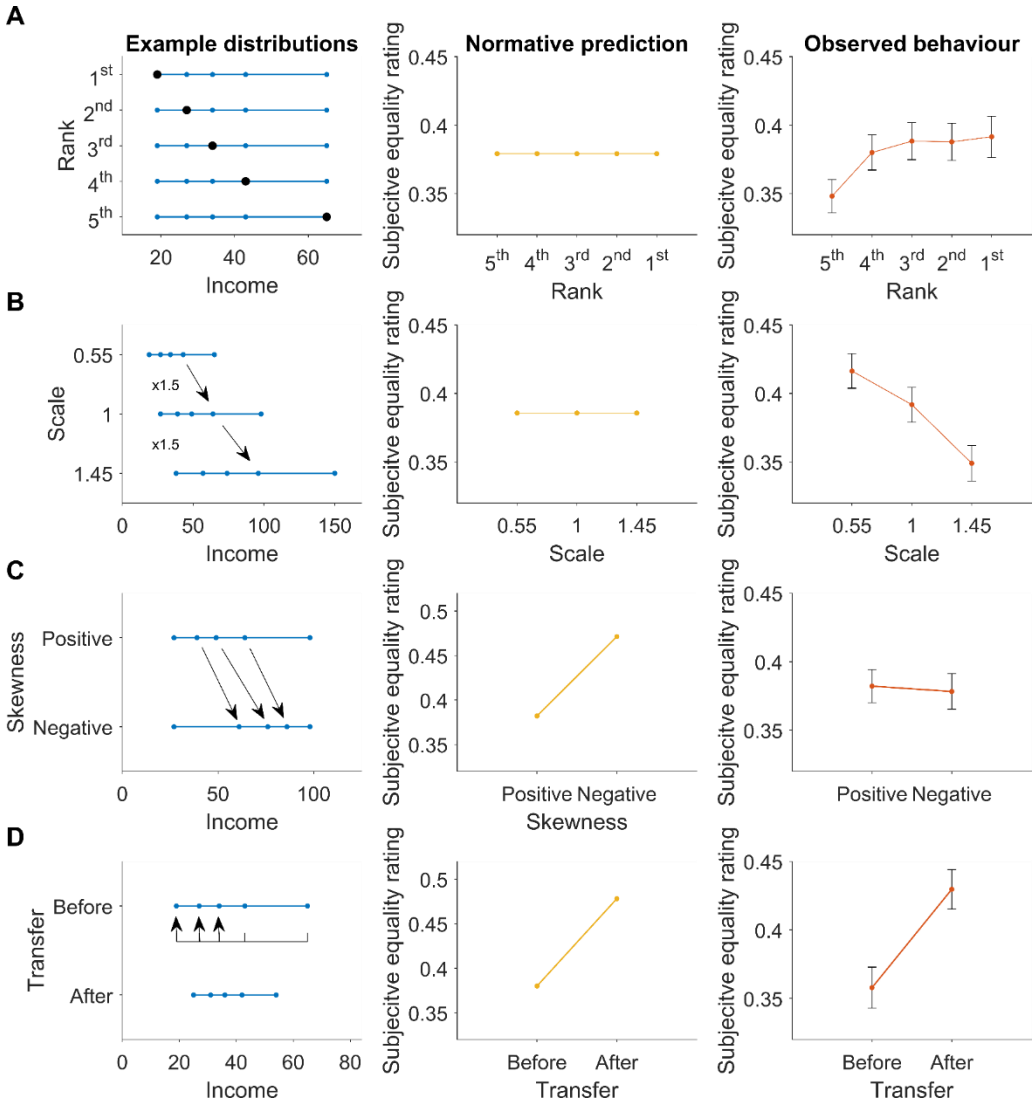


**Figure 2. Subjective perception of inequality violates normative principle underlying inequality measures.** Here we present schematically the normative principle (first column), its prediction about equality ratings (second column) and the actually observed behaviour (third column). **A**) According to anonymity principle, position of a person in the distribution (marked by black dot) should not matter for the estimation of equality. Instead, we observe that participants with lower ranks (and especially the bottom rank) judge the distributions as less equal than participants with higher ranks. **B**) According to the scale independence principle, multiplying all incomes by some factor (x1.5 in the example), should not change estimation of inequality. Instead, we observe that participants monotonically decrease their judgments of equality with increasing scale. **C**) According to the additivity principle, adding incomes to the middle incomes should increase social welfare. Increasing the mean income by doing so will also decrease inequality, as estimated by Gini coefficient. Instead, we observe that when the resulting distributions are a mirror image of each other, differing in skewness, they are judged as similarly unequal. **D**) According to the transfer principle, transferring money from richer to

poorer should increase equality. We observed that participants are consistent with the transfer principle.

**Subjective perception of inequality violates the scale-independence principle**

According to the scale-independence principle, multiplying all incomes by some factor should not change the estimation of inequality (**Figure 2B**). However, if participants focus on absolute rather than the relative magnitude of incomes, then the perceived inequality should increase proportionally to the multiplication factor. In the experiment, each level of inequality (as estimated by Gini coefficient) was presented at 3 different scales (i.e. multiplicative transformations of the same distribution), allowing us to compare the average equality judgments between these scales while holding the scale-independent inequality constant. We find that for the same Gini coefficients, distributions with higher scale were judged as more unequal than distributions with lower scales (**Figure 2B**; one-way ANOVA: $F(2, 282) = 7.47$, $p < 0.001$). We also find that a regression model that uses mean absolute difference between incomes (a scale-dependent version of the Gini coefficient, sometimes referred to as absolute Gini coefficient), fits better to subjective inequality ratings ($R^2 = 0.38$; BIC = -3607) than a regression model that uses Gini coefficient ($R^2 = 0.34$; BIC = -3277). These results suggest that subjective perception of inequality increases with the scale of the distribution, violating the scale-independence principle.

**Subjective perception of inequality violates the additivity principle.**

A related issue concerns the additivity principle and sensitivity to skewness (**Figure 2C**). According to the additivity principle, adding incomes to people whose position is not the target of aspirations for others (usually the top-income), should increase social welfare. What are the consequences of

such an addition of income? It simultaneously changes dispersion, mean income, and skewness of the distribution. Adding incomes to the middle incomes of a positively skewed distribution will always make this distribution more negatively skewed. Gini coefficient will always be lower for negatively skewed distributions than for equivalent positively skewed distributions, due to a higher mean income in the negatively skewed distributions, which acts as a scale-normalizing factor for the Gini coefficient. Based on the above, one would predict that adding incomes to the middle part of the distributions should decrease inequality.

The previous study has shown that it is not always the case: gradually adding incomes to subsequent ranks, starting from the second-highest rank and stopping at the second-lowest rank, resulted in an inverse U-shape pattern of changes in equality estimation (Amiel and Cowell, 1999). This result suggests that there is some tipping point after which adding incomes increased perceived inequality in the above case. We hypothesize this pattern occurred due to insensitivity to skewness, defined as a similar subjective perception of inequality for positively and negatively skewed distributions that are a mirror image of each other. Indeed, when we compare pairs of distributions that differed in skewness sign but matched in range of incomes and standard deviation, we find no significant difference in the subjective perception of inequality between them (**Figure 2C**; $t(94) = 0.55$, $p = 0.58$), despite a large difference in average Gini coefficient (0.37 vs. 0.21). However, it is unclear if this result can be fully explained by scale-dependence or perhaps represents a separate feature of subjective perception – an issue that we return to later while attempting to model different features of inequality perception simultaneously.

**Subjective perception of inequality is consistent with the transfer principle.**

According to the transfer principle, any transfer of income from a richer to poorer person should decrease inequality (**Figure 2D**). To test if participants are sensitive to such transfers, we compared 15 pairs of distributions that could be considered a result of such transfers. In these distributions, the sum of wealth was roughly equal, but they differed in the share of income of people at the lower positions (see **Table 1**. for statistical comparison of each pair). On average, distributions where the poor had a higher share of wealth were considered as more equal than distributions where they had a lower share (**Figure 2D**; $t(94) = -2.78$, $p < 0.01$), suggesting that subjective perceptions of inequality are consistent with the transfer principle. In all compared pairs of distributions, the average difference of perceived inequality after and before the transfer was positive, suggesting strong support for the transfer principle (**Table 1**.). Only in one case, the comparison was not significant at least at the 0.1 level, and this case involved simultaneous transfer to the poorest and the richest.

**Table 1. Test of transfer principle.** The table shows pairs of distributions that can be considered a result of the transfer of income from one to another. Some cases involve a 'leaky bucket', that is the sum of wealth after the transfer is slightly lower than before transfer. However, in none of the cases the loss is greater than 5 points. Mean equality rating between distribution D2 and D1 reports the difference between average equality ratings for each distribution. Positive values indicate an increase in inequality after the transfer. P-value is based on a non-parametric signed-rank test. * $p < 0.05$, **, $p < 0.01$, ***, $p < 0.001$.

| **Distribution D1 Before transfer** | **Distribution D2 After transfer** | **Mean equality rating D2 – D1** | **p-value** |
|---|---|---|---|
| 19 27 34 43 65 | 25 31 36 42 54 | 0.0788 | ** |
| 7 21 43 87 297 | 29 50 72 104 196 | 0.1039 | *** |
| 10 72 90 101 110 | 23 68 86 98 111 | 0.0389 | 0.072 |

| | | | |
|---|---|---|---|
| 20 79 100 113 127 | 48 77 91 103 119 | 0.0408 | * |
| 15 111 139 156 170 | 33 103 130 150 170 | 0.0655 | *** |
| 5 139 167 180 189 | 29 121 153 175 196 | 0.0207 | 0.16 |
| 13 21 31 47 94 | 25 37 43 48 54 | 0.1738 | *** |
| 10 19 30 48 110 | 22 39 46 52 59 | 0.1798 | *** |
| 8 17 29 50 129 | 19 41 50 57 65 | 0.1444 | *** |
| 7 15 28 52 155 | 17 46 57 65 73 | 0.1523 | *** |
| 5 14 27 55 189 | 15 52 65 74 82 | 0.1302 | *** |
| 13 60 76 86 94 | 15 52 65 74 82 | 0.0389 | 0.054 |
| 15 29 46 74 170 | 13 60 76 86 94 | 0.0715 | ** |
| 7 21 43 87 297 | 8 87 108 120 129 | 0.0322 | 0.053 |
| 19 27 34 43 65 | 25 31 36 42 54 | 0.0788 | ** |

A separate problem concerns the sensitivity to transfers between different parts of the distribution. Previous studies have found that although people mostly agree with the general statement that transferring money to a poorer person decreases inequality when the problem is described verbally, the results are more mixed when the problem involves comparing distributions presented numerically (Amiel and Cowell, 1999). In particular, subjective perception of inequality agrees with the transfer principle only when the money is transferred from the very rich to the poor, but not when it is transferred from poor to the poorer, suggesting greater sensitivity of inequality perception to changes in the upper part of the distribution (Amiel and Cowell, 1999). To assess more formally if this was the case in the current dataset, we fit a non-normalized Extended Gini Index - a modification of the Gini Index designed to quantify such sensitivity, equivalent to the mean absolute difference that assigns different weights to the different parts of the distribution. The fit was performed on all trials, as oppose to testing only pairs of distributions that can be considered a result of transfer of income between each other. We find that the median value of the sensitivity parameter was equal to 2.69, significantly higher than the neutral point, which for the Extended Gini Index is equal to 2 (non-parametric signed-rank test: $z = 3.38$, $p$

< 0.001). Contrary to the previous study, this result indicates that participants put more emphasis on the bottom rather than the top part of the distribution when making inequality judgments.

**Comparison of non-parametric measures approximating subjective perception of inequality**

What objective measure could approximate the subjective perception of inequality? To answer this question, we fit 12 different non-parametric inequality measures to the data, using a mixed-effects linear regression model that also included a random intercept for each participant. These included: mean, median, standard deviation, coefficient of variation, Theil index, Hoover index, Gini index, mean absolute difference, 20/20 ratio, range of incomes, top income, and bottom income (**Figure 3.**). Consistent with earlier suggestions, the mean absolute difference of incomes explained the highest amount of variance in the subjective perception of inequality ($R^2$ = 0.3801) when considering a measure that fits overall best to all participants. It is closely followed by range of incomes ($R^2$ = 0.3796) and standard deviation ($R^2$ = 0.3790).

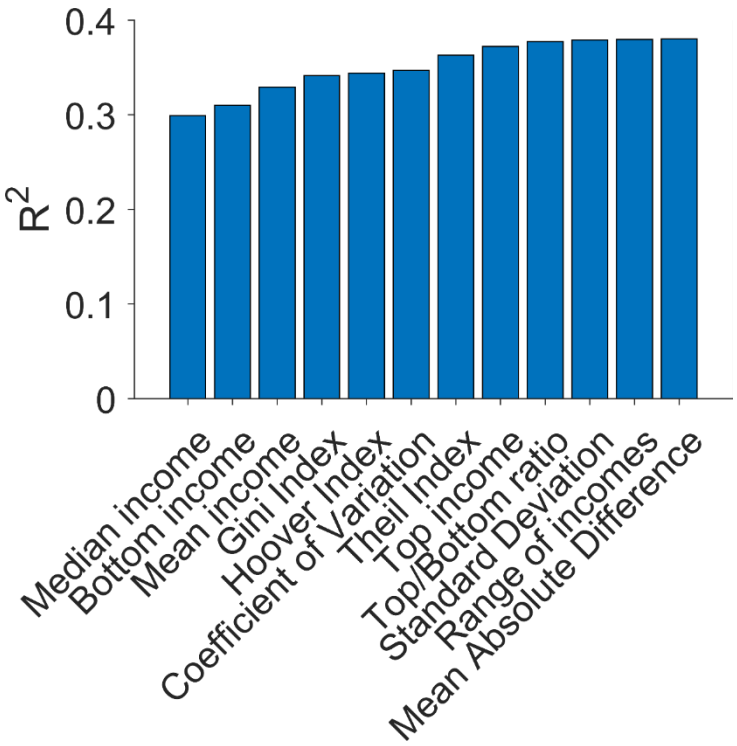


**Figure 3. Comparison of fits of non-parametric measures of inequality.**

**A new parametric measure approximating subjective perception of inequality.**

So far, we have analyzed the properties of the subjective perception of inequality in isolation from each other. However, it is possible that multiple different features can be explained by one underlying factor, as suggested before for skewness-insensitivity and scale-dependence. Furthermore, non-parametric measures do not allow for the assessment of the continuous nature of some features of the subjective perception, possibly mischaracterizing some of them. For example, participants may be scale-independent to some extent but rely on an imperfect normalization. Additionally, a measure with parameters fit individually to each person allows us to draw a picture of inter-individual variability. For these reasons, we develop a new parametric measure of subjective equality (SE) that aims to approximate the observed inequality perception.

We take the mean absolute difference (MAD) as a starting point, which is expressed as:

$$\mathrm{MAD} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|$$

Where $n$ is the number of incomes in the distribution, and $x_i$ is an individual income.

To model scale (in)dependence in a continuous way, we normalize MAD by the average income and multiply the normalization factor by the parameter $\theta \in \mathbb{R}_{\geq 0}$.

$$\mathrm{SE} = \frac{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{1 + \theta\bar{x}}$$

Where:

$$\bar{x} = \frac{1}{n}\sum x_i$$

For $\theta$ = 0, SE is equal to MAD, and for $\theta$ = 2, SE is roughly equal to the Gini index.

To account for the possible existence of skewness insensitivity mechanism separate from the scale-dependence, we used the mean of extreme incomes as an alternative normalization factor:

$$\hat{x} = \frac{1}{2}[min(x) + max(x)]$$

$$\mathrm{SE} = \frac{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{1 + \theta\hat{x}}$$

Where $\boldsymbol{x}$ is a vector of all incomes. This normalization factor can enforce scale independence while giving the same estimation of inequality for positively and negatively skewed distributions, which have the same range of incomes.

To allow for a continuous expression of skewness (in)sensitivity, we introduce parameter $\delta \in [0, 1]$.

$$\mathrm{SE} = \frac{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{1 + \theta\big((1-\delta)\bar{x} + (\delta)\hat{x}\big)}$$

For $\delta$ = 1, a person is entirely insensitive to skewness, irrespective of their scale-dependence, and for $\delta$ = 0, a person considers negatively skewed distribution as more equal, to the extent that they are scale-dependent.

To investigate if the subjective perception of inequality has a different sensitivity to transfers in different parts of the distribution, we used extended Gini index, that was developed for that purpose (Yitzhaki, 1983). Non-normalized Extended Gini (EG) is calculated as follows:

$$\mathrm{EG} = 2\sum_{i=1}^{n} w_i(x_i - \bar{x})\left[(1-F)^{v-1} - \sum w_i\,(1-F)^{v-1}\right]$$

$$F = \sum_{j=0}^{i-1} w_j + \frac{w_i}{2}$$

Where $n$ is the number of incomes in the distribution**,** $w$ is a vector of weights that will be equal to $\mathrm{w} = \frac{1}{n}$ in the case where all incomes represent equal parts of the population, and $v$ is the sensitivity parameter. For $v = 1$, a person is not sensitive to inequality, for $v < 2$ a person is more sensitive to the upper part of the distribution, for $v = 2$ they are more sensitive to the middle part of the distribution (at this parameter value EG is also equivalent to standard Gini index), and for $v > 2$ they are more sensitive to the lower part of the distribution.

Extended Gini index normalized according to the rules outlined above (or simpler normalization factors accordingly) will take the following form:

$$\mathrm{SE} = \frac{2\sum_{i=1}^{n} w_i(x_i - \bar{x})[(1-F)^{v-1} - \sum w_i\,(1-F)^{v-1}]}{1 + \theta\big((1-\delta)\bar{x} + (\delta)\hat{x}\big)}$$

Finally, in situations where the person making a judgment is also a recipient of one of the incomes, the above equation will be augmented with a person's income rank $R$.

To fit the above equation to the subjective perception of inequality that was expressed on a bounded continuous scale ranging from 0 to 1, we used the logistic function, that transforms the above continuous SE into a variable ranging from 0 to 1:

$$f(SE) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 SE + \beta_2 R)}}$$

Where $\beta_0$ is the intercept or the indifference point, $\beta_1$ and $\beta_2$ are respectively relative weights of dispersion and rank. Both *SE* and *R* are z-scored prior to being entered into the logistic function.

Overall, the full model has six free parameters: $\theta$, $\delta$, $v$, $\beta_1$, $\beta_2$, $\beta_0$, which describe scale sensitivity, skewness sensitivity, transfer sensitivity, dispersion sensitivity, rank sensitivity and indifference point respectively.

We compared the full model, with its simpler variations. As a benchmark, we also include in the comparison the best performing non-parametric measure (MAD), described in the previous section, and the two most widely used parametric measures of inequality: Atkinson index and Generalized Entropy Index. All these alternative dispersion measures were entered into the logistic regression instead of SE. To evaluate the performance of these measures, we compare them on their Bayesian Information Criterion (BIC) score, which simultaneously considers the log-likelihood of the model, and penalizes the model for its complexity, promoting a choice of the most parsimonious solutions among best-fitting models.

We find that a five-parameter version of the SE ($\theta$, $v$, $\beta_0$,$\beta_1$, $\beta_2$) that includes all parameters except a parameter that expresses skewness sensitivity in a continuous fashion (but includes skewness insensitive normalizing factor) outperformed all other models according to the BIC score (BIC = - 23775; $R^2$ = 0.52; **Table 2.**). Full model with all six parameters ($\theta$, $\delta$, $v$, $\beta_1$, $\beta_2$, $\beta_0$) explained the highest amount of variance ($R^2$ = 0.53). However, an increase of explained variance was not substantial enough to justify the inclusion of an additional parameter (BIC = -23522). These versions of the model also outperformed other widely used parametric inequality measures: the Atkinson Index (BIC = -23354; $R^2$ = 0.47) and the Generalized Entropy Index (BIC = -22722; $R^2$ = 0.42), as well as Mean Absolute Difference (BIC = -22860; $R^2$ = 0.39), which was the best non-parametric approximation of perceived inequality.

The median value of the intercept of the best fitting model suggests that participants' ratings were on average significantly below the midpoint of the scale, indicating that participants perceived the distributions as more unequal than equal ($\beta_0$ = -0.57, z = -7.07, p < 0.0001). Participants also judged the distributions as less equal the higher the quantified dispersion ($\beta_1$ = -0.51, z = -7.43, p < 0.001), supporting the idea that people are sensitive to measures of statistical dispersion, even when they are based on observation of just 5 numbers. We also find that the lower the participants rank, the more the distributions was perceived unequal, violating the anonymity principle ($\beta_2$ = 0.02, z = 3.32, p < 0.001). The median value of the scale sensitivity parameter $\theta$ was equal to 0.05 (z = -4.16, p <0.001), significantly different from 2 (which would indicate scale-independence similar to Gini Index), suggesting that people strongly violate scale-independence principle. Out of all participants, only 16.84% had a parameter value equal to or greater than 1, indicating that a great majority of participants are scale-sensitive. The median value of the transfer sensitivity parameter $v$ was equal to 2.27 (z = 4.05, p < 0.001), significantly higher than 2 (which would indicate greater emphasis on the middle income, similar to Gini index), indicating that people put a greater emphasis on the changes in incomes in the lower part of the distribution.

To test if value uncertainty could moderate any of the above effects, we compared the median parameter values between a group of participants who were informed about the value of points prior to the task, to participants who were informed about the value of points after the task. Neither of the parameters significantly differed between these two-groups ($\theta$: z = -0.00, p = 0.99; $v$: z = 1.71, p = 0.09; $\beta_0$: z =0.57, p =0.57; $\beta_1$: z = 1.09, p = 0.28; $\beta_2$: z = -0.14, p = 0.89), suggesting that uncertainty about value does not influence how people judge distributions of rewards.

**Comparison of fits of Subjective Equality Index and other measures of inequality.** The table reports the mean $R^2$ and the Bayesian Information Criterion summed over participants,

based on models fit individually to each participant. We tested different specifications of the Subjective Equality index that included a combination of four factors: scale insensitivity parameter, skewness insensitivity parameter, skewness normalization factor, and transfer insensitivity parameter. The skewness insensitivity parameter was always paired with skewness normalization factor. All models include rank of income. Models without the rank of income had a worse fit in each case (see Supplementary Table 2.) The Subjective Equality indexes are ordered from the best to the worst fitting model. The last three rows include Atkinson Index, Generalized Entropy Index and Mean Absolute Difference for comparison.

| **Scale (in)sensitivity** | **Skewness (in)sensitivity** | **Skewness normalization** | **Transfer (in)sensitivity** | **Mean $R^2$** | **BIC** |
|---|---|---|---|---|---|
| **x** | | **x** | **x** | **0.52** | **-23775** |
| x | | | x | 0.52 | -23737 |
| x | | x | | 0.48 | -23675 |
| x | x | x | x | 0.53 | -23522 |
| x | x | x | | 0.50 | -23506 |
| | | x | x | 0.49 | -23371 |
| | | | x | 0.47 | -23368 |
| | | x | | 0.43 | -23331 |
| | x | x | | 0.46 | -23146 |
| | x | x | x | 0.48 | -23126 |
| x | | | | 0.44 | -22966 |
| | | | **Atkinson Index** | 0.47 | -23354 |
| | | | **Generalized Entropy Index** | 0.42 | -22722 |
| | | | **Mean Absolute Difference** | 0.39 | -22860 |

## DISCUSSION

To understand the effect of inequality on society, we first have to decide how to measure it. Economists have suggested many different ways in which inequality can be quantified. However, we still lack the basic understanding of how people form intuitive inequality judgments based on examples of distributions of incomes presented to them. The current study fills in this gap, by characterizing in detail computations that govern subjective perceptions of statistical dispersion. We build on the seminal work of Amiel and Cowell (1999), who in a series of surveys showed that people in their judgments of pairs of

distributions tend to violate many principles assumed in economic measures of inequality. We expand this work and introduce a novel methodology. We demonstrate that people violate the anonymity principle, by being affected by their position in the distribution, the scale-independence principle, by being affected by the size of the economy, and the additivity principle, by being insensitive to the addition of incomes that transforms positively skewed distributions into negatively skewed ones. We find that people on average respect the transfer principle, according to which a transfer of income from a richer person to poorer person should decrease inequality. We synthesize these findings by developing a Subjective Equality Index – a parametric model of inequality perception that fits better to participants' data than any other conventional measure of inequality.

Amiel and Cowell (1999) have concluded that the majority of people respect the anonymity principle, agreeing that inequality of the distribution does not change if specific ranks of people in the distribution are re-arranged while holding everything else constant. In the study mentioned above, participants had to make a judgment about a hypothetical distribution of income between anonymous people. Here we take a different approach where a person making a judgment is also a recipient of one of the incomes in the distribution. In contrast to previous findings, we show that in such situations people judge the distributions as more unequal when they are at lower ranks than when they are at higher ranks.

Despite clear violation of the anonymity principle, this finding is consistent with many theories, including the Fehr-Schmidt inequality aversion model (Fehr and Schmidt, 1999) and relative deprivation theory (Hirschman, 1973), according to which disadvantageous inequality should be more aversive than advantageous inequality. We can speculate that a person's rank could exhibit its effect on general judgments of inequality through a number of different mechanisms. First of all, people could be unable to take a general perspective and ignore their personal position when evaluating the

distribution of incomes. It also possible that personal rank is a salient feature of the distribution that captures people's attention and anchors their general judgment (Markovsky, 1988). Another possibility is that one's rank induces feelings, that are subsequently interpreted as additional information about how 'good' or 'bad' is the distribution (van den Bos, 2003). Finally, when people are disadvantaged by the distribution they might be motivated to reach the conclusion that the distribution is 'bad' and therefore the wealth needs to be redistributed, but motivated to reach an opposite conclusion when they benefit from it (Kunda, 1990). As the current study does not allow us to discern the possible influences of these different mechanisms, future studies will need to employ more elaborate study designs to explain why personal rank biases general inequality judgments.

Apart from the violation of the anonymity principle, we also find strong evidence for violation of the scale-independence principle: multiplying all incomes by some factor increases perceived inequality, despite no effect on the Gini coefficient. Based on our model of subjective equality judgments, we estimate that only 16.84% of participants exhibited at least a moderate degree of scale-independence. Violation of scale-independence by subjective perceptions of inequality is shared with some 'absolute' inequality measures that do not use normalization, such as the mean absolute difference. Some economists have called for broader utilization of such absolute measures for theoretical and practical reasons (Atkinson and Brandolini, 2010; Yitzhaki, 2002; Bandyopadhyay, 2018). Closer alignment with the public perception of inequality could be another argument for wider use of such measures. Indeed, we find that the mean absolute difference fits to participants' data best, out of all non-parametric measures of inequality.

According to the additivity principle, increasing income of any person that is not the target of income aspirations (usually assumed to be the person at the top), should increase social welfare. Here we investigated if the addition of incomes to people in the middle of the distribution, that transforms a

positively skewed distribution into its negatively skewed mirror-image, would change the perception of inequality. Although according to both additivity principle and scale-independence principle inequality should decrease in such a situation, we observed that people judge distributions that are a mirror-image of each other as similarly unequal. This result could be potentially explained by full scale-dependence, as the normalization by the mean income is the main reason why most objective inequality measures would evaluate negatively skewed distribution as less equal than a positively skewed one that is a mirror-image of it. However, we do find that most participants exhibit some degree of scale-independence, and therefore should judge positively and negatively skewed distributions differently if they used mean of all incomes as a normalizing factor. Instead, our modeling results suggest that skewness insensitivity most likely originates from normalization by a skewness insensitive factor, such as the mean of extreme values. This implies that people focus on the top and bottom income, when adjusting their judgments for the scale of the economy, consistent with suggestions that these points might be the most salient parts of the distribution (Schneider, 2019; Powdthavee, Burkhauser, De Neve, 2017). In the case of positively skewed distributions, a minority or people are rich while the majority is poor, while in the case of negatively skewed distributions the majrotiy is rich, while the minority is poor. Although according to the additivity principle the latter case should be judged as the one with higher social welfare, letting the minority to stay poor while the society is getting richer might be perceived as unequal as letting the minority to get rich, as suggested by our results.

The last explored axiom is the transfer principle, according to which transferring money from a richer to poorer person should decrease inequality. In our study, most participants agreed with the transfer principle. This support is reassuring, as this axiom is fundamental for most objective inequality measures. We also explored if transfers between different parts of the distribution lead to different changes in the estimation of inequality. In line

with social welfare theories according to which inequality measures should be more concerned with the well-being of the poor than the rich (Sen, 1992), we find that participants were more sensitive to changes in the lower part of the distribution than the higher part. In other words, the transfer of money to the poor from moderately wealthy decreased inequality to a greater extent than the transfer of money to moderately wealthy from the rich. However, this pattern of results differs from the findings in the previous study (Amiel and Cowell, 1999), which suggested that people are more sensitive to changes in incomes in the upper part of the distribution. This discrepancy might be due to differences in methodology. The current study estimated the sensitivity to different parts of the distribution parametrically based on a continuous measure of perceived inequality, while the previous study drew conclusions from a general pattern of percentages of categorical responses.

The above findings are combined in a new parametric measure of Subjective Equality, which is based on the extended Gini coefficient. It consists of five parameters: a) intercept, describing an average reaction to a distribution, b) relative weight put on dispersion, describing the strength of the relation between the quantified inequality and subjective expression of inequality, c) relative weight put on person's income rank, describing to what extent a person is biased by their own position in the distribution, d) strength of normalization, regulating the extent to which the person's perception of inequality is scale-independent, and e) transfer sensitivity, describing the weight put on changes in income in different parts of the distribution. Additionally, the model includes skewness-insensitive normalizing factor, by which the measure of dispersion is divided. We show that this index of Subjective Equality fits better to participants' data than any other commonly used measure, including other parametric inequality indices such as Atkinson Index and Generalized Entropy Index, and non-parametric measures such as Gini Coefficient or Mean Absolute Difference. We also show that the five parameter version of the Subjective Equality index outperforms its simpler

specifications, and greater complexity sufficiently improves the model fit to justify the loss of parsimony. For researchers who are interested in approximating subjective perceptions of equality, but whose paradigms do not allow for the fitting of the above model, we recommend using mean absolute difference, as it was the best fitting non-parametric model in our study.

As pointed by Atkinson (1970), even seemingly objective measures of inequality are based on subjective assumptions about social welfare and arbitrary criteria of what should be important for a measure of dispersion. The subjectivity of these assumptions creates a need for an evaluation of how well they represent more widespread views on inequality. Our findings suggest that in many cases the objective measures and lay perceptions deviate from each other. In general, lay perceptions of inequality seem to be influenced by considerations of personal and societal welfare, as evident from personal rank bias, greater sensitivity to transfers to the poor than the rich, and similar evaluations of situations where a minority is much poorer than the majority and where minority is much richer than the majority. Important limitation of the current study is that subjective inequality judgments do not necessarily directly map to explicit welfare judgments, and some initial work suggests that these might differ in some cases (Amiel, Cowell, and Gaertner, 2012). Future work will need to address this problem by simultaneously asking participants about inequality, fairness and valence of evaluated distributions.

Overall, our study provides a quantitative characterization of principles governing subjective judgments of inequality and outlines a framework for future studies aiming to investigate the relationship between perceived inequality and other behaviors. In contrast to previous research (Amiel and Cowell, 1999), we do find that participants violate the anonymity principle, are almost entirely scale-dependent, and respect the transfer principle while being more sensitive to transfers in the lower part of the distribution. We also find that they are insensitive to skewness, possibly due to both scale-dependence and normalization of dispersion that is insensitive to skewness. Our Subjective

Equality Index incorporates these features into one measure while providing flexibility for estimating variability in the above patterns. In a broader context, our findings contribute to ongoing interdisciplinary efforts in characterizing the behavioral reactions to inequality and are important for many different lines of research, including numerical perception, contextual influences on valuation, and welfare economics.

## SUPPLEMENTARY MATERIAL

**Supplementary Table 1. Reward offers.** Each row corresponds to a separate trial presented to participants.

| Rank | | | | |
|---|---|---|---|---|
| **5th** | **4th** | **3rd** | **2nd** | **1st** |
| 25 | 31 | 36 | 42 | 54 |
| 22 | 29 | 35 | 42 | 59 |
| 19 | 27 | 34 | 43 | 65 |
| 17 | 25 | 33 | 44 | 73 |
| 15 | 23 | 32 | 45 | 82 |
| 13 | 21 | 31 | 47 | 94 |
| 10 | 19 | 30 | 48 | 110 |
| 8 | 17 | 29 | 50 | 129 |
| 7 | 15 | 28 | 52 | 155 |
| 5 | 14 | 27 | 55 | 189 |
| 34 | 44 | 51 | 60 | 79 |
| 30 | 41 | 50 | 62 | 88 |
| 27 | 39 | 49 | 64 | 98 |
| 23 | 36 | 48 | 66 | 111 |
| 20 | 34 | 47 | 68 | 127 |
| 17 | 31 | 46 | 71 | 146 |
| 15 | 29 | 46 | 74 | 170 |
| 12 | 26 | 45 | 77 | 201 |
| 10 | 24 | 44 | 81 | 242 |
| 7 | 21 | 43 | 87 | 297 |
| 48 | 64 | 76 | 90 | 119 |
| 43 | 60 | 75 | 93 | 133 |
| 38 | 57 | 74 | 96 | 150 |
| 33 | 53 | 73 | 100 | 170 |
| 29 | 50 | 72 | 104 | 196 |
| 25 | 47 | 71 | 109 | 226 |
| 21 | 43 | 70 | 114 | 265 |
| 18 | 40 | 69 | 120 | 315 |
| 15 | 37 | 68 | 127 | 379 |

| | | | | |
|---|---|---|---|---|
| 12 | 33 | 67 | 136 | 466 |
| 25 | 37 | 43 | 48 | 54 |
| 22 | 39 | 46 | 52 | 59 |
| 19 | 41 | 50 | 57 | 65 |
| 17 | 46 | 57 | 65 | 73 |
| 15 | 52 | 65 | 74 | 82 |
| 13 | 60 | 76 | 86 | 94 |
| 10 | 72 | 90 | 101 | 110 |
| 8 | 87 | 108 | 120 | 129 |
| 7 | 110 | 134 | 147 | 155 |
| 5 | 139 | 167 | 180 | 189 |
| 34 | 53 | 62 | 69 | 79 |
| 30 | 56 | 68 | 77 | 88 |
| 27 | 61 | 76 | 86 | 98 |
| 23 | 68 | 86 | 98 | 111 |
| 20 | 79 | 100 | 113 | 127 |
| 17 | 92 | 117 | 132 | 146 |
| 15 | 111 | 139 | 156 | 170 |
| 12 | 136 | 168 | 187 | 201 |
| 10 | 171 | 208 | 228 | 242 |
| 7 | 217 | 261 | 283 | 297 |
| 48 | 77 | 91 | 103 | 119 |
| 43 | 83 | 101 | 116 | 133 |
| 38 | 92 | 114 | 131 | 150 |
| 33 | 103 | 130 | 150 | 170 |
| 29 | 121 | 153 | 175 | 196 |
| 25 | 142 | 180 | 204 | 226 |
| 21 | 172 | 216 | 243 | 265 |
| 18 | 213 | 264 | 293 | 315 |
| 15 | 267 | 326 | 357 | 379 |
| 12 | 342 | 411 | 445 | 466 |

**Supplementary Table 2. Comparison of fits of Subjective Equality Index and other measures of inequality, without including the rank of income.**

| Scale (in)sensitivity | Skewness (in)sensitivity | Skewness normalization | Transfer (in)sensitivity | Mean $R^2$ | BIC |
|---|---|---|---|---|---|
| **x** | | **x** | **x** | **0.48** | **-23696** |
| x | | | x | 0.48 | -23633 |
| x | | x | | 0.44 | -23608 |
| x | x | x | x | 0.49 | -23421 |
| x | x | x | | 0.46 | -23431 |
| | | x | x | 0.43 | -22917 |

| | | | | | |
|---|---|---|---|---|---|
| | | | x | 0.39 | -23278 |
| | | x | | 0.43 | -23302 |
| | x | x | | 0.41 | -23084 |
| | x | x | x | 0.44 | -23054 |
| x | | | | 0.40 | -22515 |
| | | | **Atkinson Index** | 0.43 | -23276 |
| | | | **Generalized Entropy Index** | 0.38 | -22654 |
| | | | **Mean Absolute Difference** | 0.35 | -22814 |

## REFERENCES


Amiel, Y., and Cowell, F. (1999). *Thinking about Inequality: Personal Judgment and Income Distributions*. Cambridge University Press.

Atkinson, A. B. B., Andrea. (2010). On Analyzing the World Distribution of Income. *World Bank Economic Review*, *24*(1), 1–37. https://doi.org/10.1093/wber/lhp020

Atkinson, A. B. (1970). On the measurement of inequality. *Journal of Economic Theory*, *2*(3), 244–263. https://doi.org/10.1016/0022-0531(70)90039-6

Bandyopadhyay, S. (2018a). The absolute Gini is a more reliable measure of inequality for time dependent analyses (compared with the relative Gini). *Economics Letters*, *162*, 135–139. https://doi.org/10.1016/j.econlet.2017.07.012

Blackorby, C., Bossert, W., and Donaldson, D. (1999). Income Inequality Measurement: The Normative Approach. In J. Silber (Ed.), *Handbook of Income Inequality Measurement* (pp. 133–161). Springer Netherlands. https://doi.org/10.1007/978-94-011-4413-1_4

Cruces, G., Perez-Truglia, R., & Tetaz, M. (2013). Biased perceptions of income distribution and preferences for redistribution: Evidence from a survey experiment. *Journal of Public Economics*, 98, 100–112. https://doi.org/10.1016/j.jpubeco.2012.10.009

Dalton, H. (1920). The Measurement of the Inequality of Incomes. *The Economic Journal*, 30(119), 348–361. JSTOR. https://doi.org/10.2307/2223525

Fehr, E., and Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. JSTOR.

Hirschman, A. O. (1973). The changing tolerance for income inequality in the course of economic development. *World Development*, *1*(12), 29–36. https://doi.org/10.1016/0305-750X(73)90109-5

Kahneman, D., and Thaler, R. H. (2006). Anomalies: Utility Maximization and Experienced Utility. *Journal of Economic Perspectives*, *20*(1), 221–234. https://doi.org/10.1257/089533006776526076

Kiatpongsan, S., and Norton, M. I. (2014). How Much (More) Should CEOs Make? A Universal Desire for More Equal Pay: *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691614549773

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Lazear, E. P., and Shaw, K. L. (2009). *The Structure of Wages: An International Comparison*. University of Chicago Press.

Mantonakis, A., Rodero, P., Lesschaeve, I., and Hastie, R. (2009). Order in Choice: Effects of Serial Position on Preferences. *Psychological Science*, *20*(11), 1309–1312. https://doi.org/10.1111/j.1467-9280.2009.02453.x

Markovsky, B. (1988). Anchoring Justice. Social Psychology Quarterly, 51(3), 213–224. JSTOR. https://doi.org/10.2307/2786920

Niehues, J. (2014). *Subjective Perceptions of Inequality and Redistributive Preferences: An International Comparison*.

Norton, M. I., and Ariely, D. (2011). Building a Better America—One Wealth Quintile at a Time: *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691610393524

Pinkovskiy, M., and Sala-i-Martin, X. (2009). *Parametric Estimations of the World Distribution of Income* (No. 15433; NBER Working Papers). National Bureau of Economic Research, Inc. https://ideas.repec.org/p/nbr/nberwo/15433.html

Powdthavee, N., Burkhauser, R. V., & De Neve, J.-E. (2017). Top incomes and human well-being: Evidence from the Gallup World Poll. Journal of Economic Psychology, 62, 246–257. https://doi.org/10.1016/j.joep.2017.07.006

Read, D., Antonides, G., van den Ouden, L., and Trienekens, H. (2001). Which Is Better: Simultaneous or Sequential Choice? *Organizational Behavior and Human Decision Processes*, *84*(1), 54–70. https://doi.org/10.1006/obhd.2000.2917

Schneider, S. M. (2019). Why Income Inequality Is Dissatisfying—Perceptions of Social Status and the Inequality-Satisfaction Link in Europe. European Sociological Review, 35(3), 409–430. https://doi.org/10.1093/esr/jcz003

Sen, A., and Foundation, R. S. (1995). *Inequality Reexamined*. Harvard University Press.

Temkin, L. S. (1993). *Inequality*. Oxford University Press.

van den Bos, K. (2003). On the Subjective Quality of Social Justice: The Role of Affect as Information in the Psychology of Justice Judgments. Journal of Personality and Social Psychology, 85(3), 482–498. https://doi.org/10.1037/0022-3514.85.3.482

Yitzhaki, S. (1983). On an Extension of the Gini Inequality Index. *International Economic Review*, *24*(3), 617–628. JSTOR. https://doi.org/10.2307/2648789

Yitzhaki, S. (2003). Gini's Mean difference: A superior measure of variability for non-normal distributions. *Metron - International Journal of Statistics*, *LXI*(2), 285–316.

# Chapter 4

# When the grass is greener on the other side of the border: how the wealth of foreign countries affects our well-being.

**Filip Gesiarz[*1], Jan-Emmanuel De Neve[2], Tali Sharot[*1]**

[1]Affective Brain Lab, Department of Experimental Psychology, University College London, London, UK

[2]Saïd Business School, University of Oxford, Oxford, UK

* to whom correspondence should be addressed: filip.gesiarz.15@ucl.ac.uk

**ABSTRACT**

Do personal life evaluations change depending on relative living standard of people in other foreign countries? To investigate this possibility, we quantified the relative position of countries and individuals within their regions of the world and tested if they are related to ratings of the current well-being. We used responses from the World Gallup Poll – a representative survey of over 2 million individuals from 154 countries. We show that an individual's life satisfaction is not only related to the relative position of their living standard in comparison to people in other countries, but is also related to a relative position of their country in international rankings of living standards, suggesting a coexistence of international comparisons based on personal and national identity. Independently, we find that inequality of living standards between countries is negatively related to average well-being across different regions of the world. A model that incorporated country's international rank, as well as inequality between countries in different regions of the world, outperformed 13 different models based on previous theories describing how relative value could influence well-being. These findings suggest that social comparisons affecting well-being extend beyond national borders, and span from comparisons of personal living standard with the situation of people in other countries, to more general considerations of international inequality.

## INTRODUCTION

Does well-being of a country depend on the living standards of its neighboring countries? The standard approach to analyzing factors underlying the average life-satisfaction in a country is to look at factors directly affecting citizens within a country, such as quality of healthcare, access to education, or living wages adjusted for costs of living (Anand and Sen, 1994; Helliwell, Huan, Wang, 2019). Many studies have also suggested that having a relatively better living standard than a comparison group matters at least as much for life satisfaction as the actual living standard: it has been shown that people are more satisfied with their lives when they are more wealthy than their neighbors vs. less wealthy than their neighbors, irrespective of their actual wealth (Firebaugh and Schroeder, 2009; Luttmer, 2005; Brown, Gray, and Roberts, 2019). Studies that demonstrated the effect of relative wealth have predominantly focused on comparisons within and between peer-groups, workplaces or districts. However, globally the most significant differences in living standards are not between different occupations, demographics, or communities, but between people living in different countries: according to some estimates country of birth accounts for as much as 66% of the variation in living standards worldwide (Milanovic, 2015). Despite recognition that relative comparisons matter, the potential effect of an international context on individuals' well-being remains relatively unexplored.

Past work on inter-group relations suggests that people often identify themselves with the group that they belong to and incorporate the relative position of their group into their sense of self-worth (Ellemers, Knippenberg and Wilke, 1990; Smith and Tyler, 1997). Nationality is arguably one of the most important group-defining characteristics. Therefore, it is possible that when people evaluate their lives, they not only think about their living standard but also about the living standard of their fellow compatriots relative to the living standard of people in other countries. Here we investigate if a person's well-being could be affected by such international comparisons.

We suggest that individuals engage in at least two different types of international comparisons: country-level and personal-level. Country-level comparisons would be related to a status of a country on an international stage, based on the average wealth of its citizens relative to average wealth of citizens in other countries. Such contrast could be related to group comparisons based on national-identity, similar to comparisons that have been observed for other group defining characteristics, such as gender or race, and which have been shown to be related to feelings of pride and shame (Smith and Tyler, 1997; Salice and Sánchez, 2016). For example, a person could derive some utility from the knowledge that their country is doing well on the international stage, irrespective of their personal wealth. On the other hand, personal-level international comparisons would involve a contrast of an individual's own living standard with the living standard of people in other countries, which could be related to counterfactual thoughts about the life that the person could live if they moved to a different country. For example, a knowledge that a person would have a better living standard if they lived in a different country, despite a high position in their own country, could have a negative effect on a person's well-being. The possibility of such comparisons assumes that people have at least some rudimentary knowledge about the living standards in other countries. In line with this assumption, it has been found that subjective perceptions of living standards of foreign countries correlate with Gross Domestic Product of these countries, making it likely that people might use such information to position themselves and their country in the international context (Lahusen and Kiess, 2016; Delhey and Kohler 2006).

One methodological challenge concerning studying the potential effect of international comparisons is the lack of consensus about what aspects of the distribution people focus on when comparing themselves to others in a multi-agent setting. Different theories and experiments suggest that people might be affected by comparison to the mean income (Helson, 1964; Ferrer-i-Carbonell, 2005), top income (Powdthavee et al., 2017), their income's rank

(Brown et al., 2008), position in the range of incomes (Soltani, De Martino, and Camerer, 2012; Hagerty, 2000; Rangel and Clithero, 2012), inequality between incomes (Oishi, Kesebir and Diener, 2011), disadvantageous and advantageous inequality (Rutledge et al., 2016; Fehr and Schmidt, 1999), or some combination of these ideas (Powdthavee et al., 2017; Gesiarz, De Neve, Sharot, 2020; Clark and D'Ambrosio, 2015). Here we pit against each other 13 different operationalizations of how international context of others' incomes could affect a person's well-being, allowing us not only to document the existence of the phenomenon but also pinpoint the most plausible mechanisms that are involved in international comparisons.

The proposed hypotheses are tested using the Gallup World Poll - a representative survey covering 98% of world's countries, including responses from more than 2.08 million individuals gathered annually during the years 2006 - 2018. We focus on evaluative well-being, defined as satisfaction about one's life overall, and measured by a Cantril Ladder. Our findings provide an extension of the relative income hypothesis to an international scale, putting it to a test in a global sample. By evaluating predictions of different theories of social comparisons we advance several theoretical avenues and broaden our understanding of the impact of globalization on human well-being.

## RESULTS

Following the findings that people tend to compare themselves to similar or proximate others (Clark and Senik, 2010; Pérez-Asenjo, 2011), we defined international comparison groups based on the regions of the world (according to the regional division made in the Gallup World Poll; see **Figure 2A**). These regions overlap with cultural clusters identified by other authors (Ronen and Shenkar, 2013). An alternative approach would involve defining comparison groups based on sharing a border. However, such an approach ignores broader cultural, political and historical ties between countries (e.g. the former Soviet Union) and is problematic in the case of countries separated by

bodies of water (e.g. the United Kingdom). After defining the comparisons groups, we quantified the international comparisons within these groups as described below.

Higher rank of a country and lower inequality between countries is associated with higher average well-being, irrespective of the absolute living standard of a country.

We perform the analysis on two levels: country-level, using representative samples, and on an individual level, taking into account other factors that might underlie the effect of international comparisons, but using a non-representative sample due to missing values for some control variables. We first focus on average life satisfaction in a representative sample of individuals in 154 countries. Previous research has used many different conceptualizations of social comparisons. To establish what aspects of the distribution of living standards matter most for the international comparisons, we run a series of regression models with thirteen different quantifications of relative value (see Methods and **Figure 1.**, for the full list). Apart from the relative value, all models included also the absolute living standard of a country. To select the best fitting and most parsimonious model, we used the Bayesian Information Criterion (BIC), which uses information about log-likelihood to assess the model fit, and penalizes the model for its complexity (Wasserman, 2000). Based on this criterion, we find that a model incorporating the rank of a country and inequality between countries in the regions of the world best accounts for the differences in life satisfaction observed in the data (See **Figure 1.**).

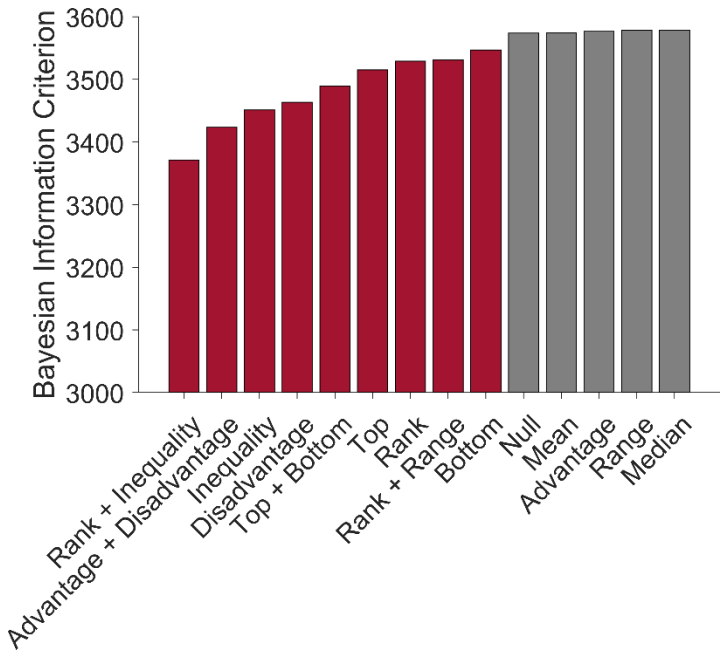


**Figure 1. Models of living standards comparisons.** The plot shows the fits of different models of international comparisons. Each model comprised of the international comparison component and absolute living standard of a country (log gross national income per capita at purchasing power parity). To compare the models, we used the Bayesian Information Criterion. Red bars indicate which models were better than the Null model consisting of only absolute living standards. Labels indicate as follows: (a) median: difference from the median living standard in the region, (b) mean: difference with the mean living standard in the region, (c) top: difference from the highest living standard in the region, (d) bottom: difference from the lowest living standard in the region, (e) range: living standard normalized by the range of living standards in the region, (f) rank: rank of living standard in the region, (g) disadvantage: disadvantageous inequality, (h) advantage: advantageous inequality, (i) inequality: inequality of living standards between countries. The best model according to the BIC score was a model consisting of a country's rank in the region of the world and inequality between countries in the region of the world.

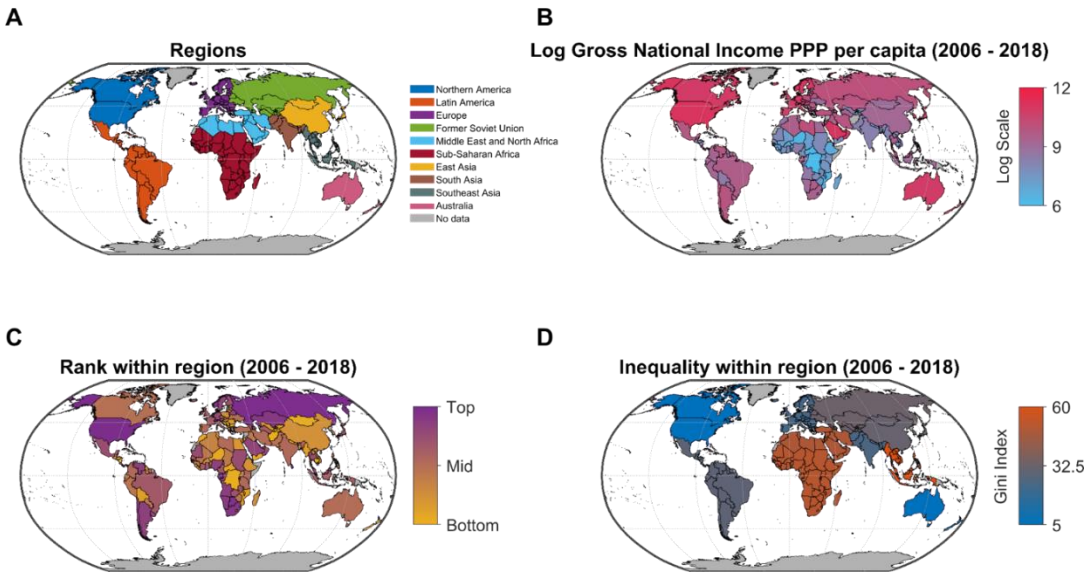


**Figure 2. Rank of a country and inequality between countries of living standards in the regions of the world.** The plot illustrates the distribution of variables used in the rank-inequality model of international comparisons, averaged over the years 2006 -2018. (**A**) Shows regions of the World, based on the classification made by Gallup World Poll. (**B**) Shows absolute living standards of countries, based on Log Gross National Income at Purchasing Power Parity per capita (Log GNI PPP pc). (**C**) Shows country's rank of living standards in the region of the world, normalized within each region to range from 0 to 1. (**D**) Shows inequality of living standards between countries in each region, quantified as Gini index.

We find that absolute living standard ($\beta = 0.69$, $p < 0.0001$) and country's rank in a region of the world ($\beta = 0.12$, $p < 0.0001$) are positively related to average life satisfaction, while the inequality between countries in the regions of the world ($\beta = -0.28$, $p < 0.0001$) is negatively related to average life satisfaction (See **Figure 2.** for the distribution of variables used in the winning model). In other words, controlling for the actual living standard in a country, countries with a relatively higher living standard than others in the region have higher average life satisfaction than countries with a lower living standard. Independently, countries in regions of the world with more unequal living standards between countries have lower average life satisfaction than countries in the regions with similar living standards. These results suggest a world-wide and population-wide effect of international comparisons on average life satisfaction of countries.

An individual's life satisfaction is related to personal rank, rank of their country and inequality between countries in the regions of the world.

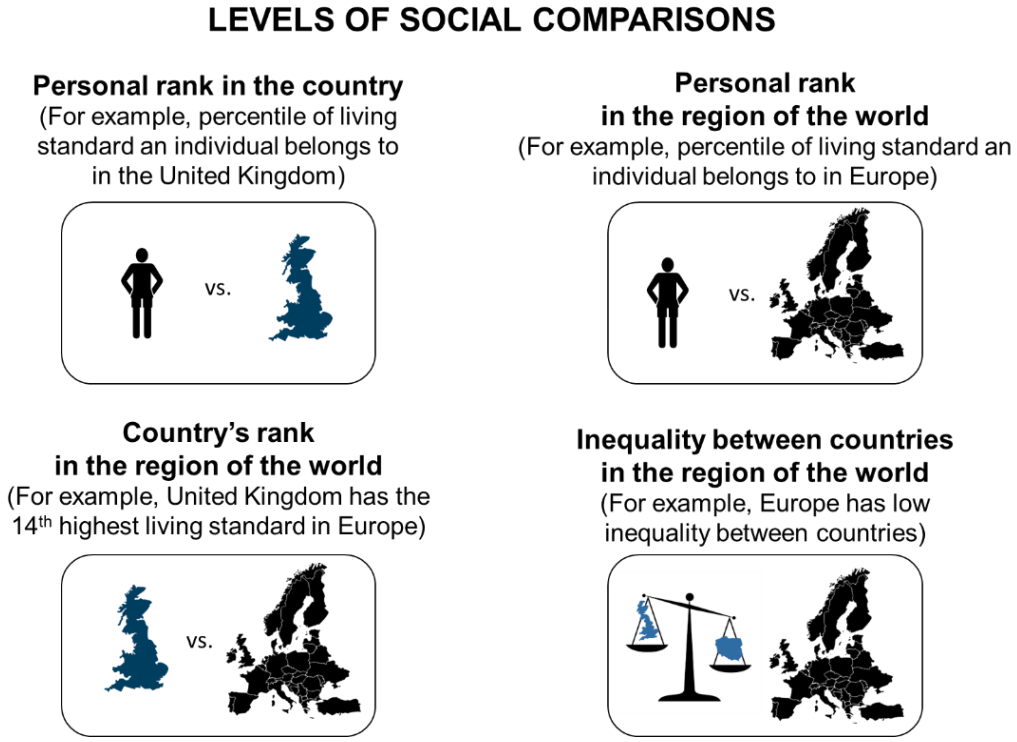


**Figure 3.** The panel illustrates different levels of social comparisons that were included in the regression models. At the country level, a person can focus on the rank of their income in their own country. At the international level, they can engage in at least three different types of comparisons characterized by a different extent of generality. At the lowest level, they can focus on their personal rank in their region of the world, e.g. in Europe. At the mid-level, they can focus on the rank of their country in the region of the world that requires an abstraction from a personal situation. At the highest level, they can focus on inequality between countries in the region of the world that requires abstraction from a situation of their own country.

The above analysis focused on aggregated life satisfaction in representative samples in 154 countries. To check the robustness of the effect, we next investigate if the above results hold when predicting the life satisfaction of individual respondents while controlling for a range of variables that might underlie the effect of the international rank of a country and inequality between countries in the regions of the world. Analysis of individuals additionally allows us to differentiate personal-level from country-level international comparisons by including rank of a person in their country, and

rank of a person in their region of the world - both calculated as income percentile that a person belongs to in the population (See Figure 3. for the classification of different levels of social comparisons and Methods section for details).

We control for three types of variables: (i) characteristics of an individual (e.g., household income), (ii) characteristics of a country that an individual is living in (e.g., country's unemployment rate), (iii) characteristics of a region of the world that an individual is living in (e.g., political stability of the region). Additionally, the model included a fixed effect for the year of the survey, country and region of the world, that allowed us to control for the effects specific for years, countries or regions not accounted for by other variables (see **Figure 5.** for the full list of the control variables). Consistently with the results about the average life satisfaction on a country level, we find that rank of a country in a region of the world is related positively ($\beta = 0.09$, $p < 0.01$), and inequality between countries is related negatively ($\beta = -0.51$, $p < 0.0001$) to life satisfaction of an individual (for full results, see Supplementary Table 1.). Additionally, we find independent positive effects of a personal rank in a country ($\beta = 0.27$, $p < 0.0001$) and personal rank in a region of the world ($\beta = 0.16$, $p < 0.0001$), suggesting a coexistence of multiple levels of social comparisons, with different degrees of generality (See **Figure 3.**). These results are robust to several different specifications of the regression models (see Supplementary Material), including a sequential regression approach that decorrelates the absolute living standard and international comparisons.

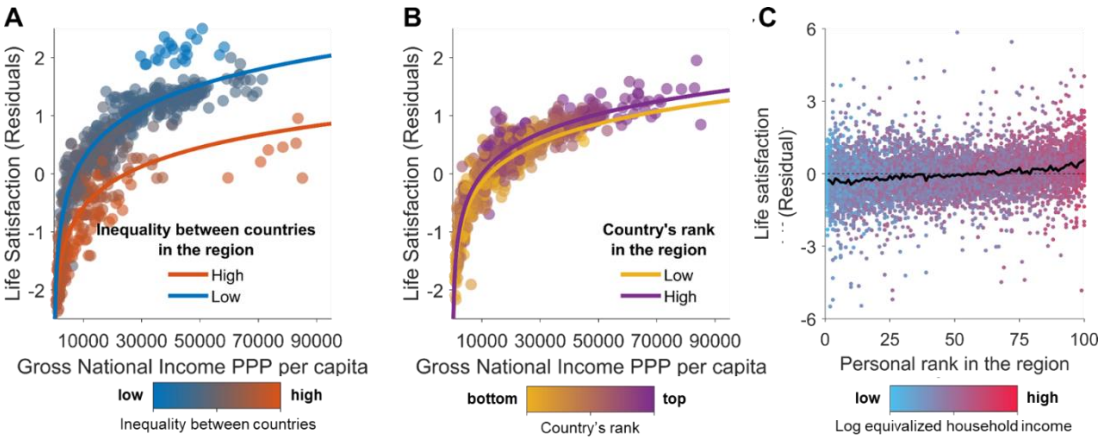

**Figure 4. An individual's life satisfaction is related to the personal rank in the region of the world, rank of their country, and inequality between countries.** The plot illustrates the effects of international comparisons and absolute living standards. In panels (A) and (B) each point represents an average residual life satisfaction from the model for a given year and country when (**A**) influence of absolute living standards in the country and inequality set to 0 (mean value); (**B**) influence of absolute living standards in the country and country's rank are set to 0 (mean value). Lines in (**A**) and (**B**) represent the best fitting log function to points divided by a median split. Colors in (**A**) represent the value of inequality between countries and in (**B**) the country's rank in the region of the world on a continuous scale. In panel (**C**) each point represents an average residual life satisfaction from the model for a given percentile in the region of the world, separately for each country, when the influence of rank in the region and personal living standard is set to 0 (mean value). Blackline represents the average value for each percentile. The colors represent the value of a person's absolute living standard on a continuous scale.

To evaluate how much international comparisons matter with respect to other variables we conduct relative importance analysis (LeBreton and Tonidandel, 2008) – a method that allows assessing proportionate contribution of each predictor to $R^2$ of the model, considering both the unique contribution of each predictor by itself and its contribution when combined with the other predictors. Overall our model explains 31.35% of the variance in the life satisfaction of individuals. 6.11% of this explained variance can be attributed to personal rank in the country, 6.09% to personal rank in the region of the world, 1.91% to country's rank in the region of the world, and 1.76% to inequality between countries in the region of the world(see **Figure 5.** for comparisons with other variables). Overall, 9.76% of the explained variance can be attributed to international comparisons. For a reference, the absolute living standard in the country and personal living standard account for the 2.47% and 1.62% of the explained variance, respectively.

## DISCUSSION

As early as in the 1960s scientists have noticed that the world is becoming a "global village", where advances in technology bring people closer together and blur traditional national boundaries (McLuhan, 1964). In this study, we explore the implications of this situation on human well-being. Specifically, we ask if the human tendency to compare oneself to others extends beyond national borders, and to what degree this affects how people evaluate their lives. To that end, we used data from the Gallup World Poll, which included a representative sample from 154 countries, constituting 98% of all countries in the world. We quantified relative living standards in each of these countries in comparison to other countries. Our results suggest that there is a significant relationship between international relative living standard and the average life satisfaction in a country as well as the life satisfaction of individuals. We find that these effects cannot be explained by differences between absolute living standards or other control variables.

Our findings uncover a multi-level structure of social comparisons. We show that life satisfaction is not only related to relative comparisons with people in one's own country, but also with relative comparisons to people in other countries, and comparisons of one's country to other countries. The distinction between personal and country-level international comparisons is an important one, as it highlights different processes that might be affecting well-being. Personal-level international comparisons are most likely an extension of comparisons made within-country, such as those between neighbours or co-workers. Country-level international comparisons are qualitatively different, as they require abstraction from the personal situation and identification with a larger group. For example, a person might feel bad because the group they belong to has a low living standard, irrespective of their personal absolute living standard being high or low. This distinction is directly related to personal and social identity (Luhtanen and Crocker, 1992), or egoistic and fraternalistic orientation constructs (Osborne et al., 2015)

identified in social identity and relative deprivation theories, respectively. These terms have been used to explain patterns of identification with groups (Mussweiler, Gabriel and Bodenhausen, 2000), collective emotions (Campo, Mackie, and Sanchez, 2019), reactions to discrimination (Eccleston and Major, 2006) and inter-group relations (Ramiah, Hewstone, and Schmid, 2011), but to the best of our knowledge have not been studied in the context of well-being and international comparisons.

The idea that social comparisons can extend beyond national borders has been suggested before in the context of the expansion of the European Union. It has been theorized that the reference groups to which Europeans compare themselves to will broaden in scope with advances in European integration (Whelan and Maitre, 2009a; Whelan and Maitre, 2009b). Indeed, one study has shown that more people in Germany than in Hungary are able to assess the living standards of eight other European nations, presumably due to the latter being a newer member of European Union; and that these pair-wise comparisons between living standards of other European nations can have a negative effect on one's life satisfaction if being disadvantageous (Delhey and Kohler, 2006). Another study found that individuals in former post-soviet countries, namely Poland, Ukraine, Hungary, and Georgia, identify western European countries as the most critical comparison group almost as frequently as they compare themselves to the wealthy citizens in their own country; moreover, this tendency was associated with lower life satisfaction (Sági, 2011). The effect of international comparisons has also been suggested in a study of the well-being of 15 western European countries that were part of the union at least since 1995. In particular, it has been shown that sharing a border with a more prosperous country has a negative effect on the country's average life satisfaction, even though all of the investigated countries in this study represented high-income countries (Becchetti et al., 2013). Our findings provide robust evidence for the existence of multiple types of international comparisons that are not limited to Europe but are prevalent globally,

suggesting that history of international integration is not necessary for the effect of the international comparison to occur. We further pinpoint the most plausible ways in which people might compare themselves to people in other countries.

To identify the most important aspects of the international distribution of wealth for international comparisons, we took a data-driven approach, pitting the predictive power of 12 different models of social comparisons against each other. The models included comparisons with the richest country, comparison with the poorest country, comparisons with the mean or median international living standard, disadvantageous or advantageous inequality of a country, rank of a country, inequality between countries, position of the country in the range between the poorest and the richest country, or a combination of these factors described previously in the literature. The model consisting of the rank of a country and inequality between countries outperformed all other alternatives. This result is in line with earlier suggestions that reactions to inequality can be broken down to two components: relative comparison of one's situation to others and general evaluation of income disparities in the group (Clark and D'Ambrosio, 2015). The importance of rank is consistent with studies showing that rank of income has a more significant impact on well-being than income itself (Boyce, Brown and Moore, 2010), as well as studies suggesting that people naturally engage in ordinal rather than absolute comparisons (Stewart, Chater, and Brown, 2006). The role of inequality is consistent with studies showing its impact on well-being in cross-national surveys (Powdthavee et al., 2017; Oishi, Kesebir, and Diener, 2011) and laboratory experiments with small groups (Rutledge et al., 2016; Gesiarz, De Neve, Sharot, 2020). Simultaneous importance of inequality and rank for well-being has also been demonstrated before in laboratory settings (Gesiarz, De Neve, Sharot, 2020).

One explanation for the negative effect of international inequality on well-being could be that historical factors that lead to inequality in the region

in the first place could also be related to the competition between countries or other variables promoting inter- and intra-national tensions. However, our analysis included proxy controls for many of these potential factors, such as the value of military expenses and foreign investment, indexes of political instability, freedom, and globalization, or numbers of migrants and refugees; limiting the extent to which such factors could explain the observed effect, and suggesting a more direct relationship between well-being and international inequality.

Our findings provide a framework for considering the impact of international context on the well-being of nations. We uncover the effects of multi-level international comparisons that people engage in, involving comparisons based on both personal and country-level identities. Considerations of country's rank and inequality between countries were the two most important factors through which international comparisons affected well-being. The effects of international comparisons were moderated by freedom of movement, suggesting that differences in opportunities might be one of the reasons why such comparisons matter for people. In a broader context, this framework provides a new angle for research that tries to explain why citizens of some countries might be dissatisfied with their lives despite having high absolute living standards, or why some regions of the world are unhappier than others.

## METHODS

All data analysis was performed using MATLAB 2019a and R 3.6.1 software.

*Well-being data.*

The Gallup's World Poll surveyed residents from 154 countries, using randomly selected and nationally representative samples. The representative samples have the same distribution of age, gender, education and socioeconomic status as the national population. To ensure national

representativeness of the samples, Gallup World Poll assigns a weight to each individual, correcting for oversampling or under-sampling from some demographic groups. These weights were used in all our regression models. The analysis covers the years from 2006 to 2018. For each year and each country, the survey included around 1000 individuals (although not all countries were included in the poll each year).

Evaluative well-being (thereafter called life-satisfaction) was measured by Cantril Ladder question in the Gallup World Poll, in which the respondent had to position themselves on a ladder, where 0 represented the worst possible life and 10 represented the best possible life.

*Definition of living standards.*

We use a proxy for absolute living standards in a country based on Log Gross National Income at Purchasing Power Parity per capita in international dollars (Log GNI PPP pc, thereafter referred to as the living standard) - a measure superior, according to some authors, in quantifying living standards to Gross Domestic Product (Capelli and Vaggi, 2013). A proxy for the absolute living standard of an individual was based on equivalized log household income. Equivalization takes into account the fact that costs of living scale non-linearly with household size and was achieved by dividing total household income by square root of number of household members, following the methodology in OECD reports (2011).

*Comparison groups.*

Based on the previous studies showing that people tend to compare themselves to a reference groups with similar characteristics (Clark and Senik, 2010; Pérez-Asenjo, 2011), we created comparisons groups based on regions identified by the Gallup World Poll: Northern America, Latin America, Europe, Former Soviet Union, Middle East, and North Africa, Sub-Saharan Africa, East Asia, South Asia, Southeast Asia, and Australia.

*Country-level analysis.*

The country-level analysis was performed on life satisfaction averages calculated for each country and each year, based on weighted responses ensuring the representativeness of the sample. These averages were based on a total of 2.082971 million responses. To assess the relationship between international comparisons and well-being we used regression analysis, which included absolute living standards in the country for each year and variables describing the effect of international comparisons. Comparisons of living standards can be made based on many different criteria. To not limit our findings to a particular conceptualization of social comparisons, we tested a range of different possible ways in which citizens of one country can compare their living standards to living standards in other countries. These included the following: (a) difference from the median living standard in the region, (b) difference from the mean living standard in the region (Helson, 1964; Ferrer-i-Carbonell, 2005), (c) difference from the highest living standard in the region (Powdthavee et al., 2017), (d) difference from the lowest living standard in the region (Kuziemko et al., 2014), (e) living standard normalized by the range of living standards in the region (Soltani, De Martino, and Camerer, 2012; Hagerty, 2000; Rangel and Clithero, 2012), (f) rank of living standard in the region (Brown et al., 2008), (g) advantageous inequality of living standards (Rutledge et al., 2016; Fehr and Schmidt, 1999), (h) disadvantageous inequality of living standards (Rutledge et al., 2016; Fehr and Schmidt, 1999), (i) inequality of living standards between countries in the region (Oishi, Kesebir and Diener, 2011). Based on conceptualizations of social comparisons in previous studies we also included: (j) living standard normalized by the range of living standards in the region and rank of a living standard in the region (Parducci, 1995), (k) advantageous and disadvantageous inequality (Fehr and Schmidt, 1999), (l) rank of a living standard in the region and inequality in the region (Gesiarz, De Neve, Sharot, 2020; Clark and D'Ambrosio, 2015).

The rank of a country was computed as a normalized value ranging from 0 to 1, for the lowest and highest rank in the region, respectively:

$$Rank = \frac{i-1}{n-1}$$

Where *i* is the living standard index in a set of living standards ordered from lowest to highest and *n* is the number of countries in the region.

Range-normalized living standard was computed as follows:

$$Range = \frac{x_i - x_{bottom}}{x_{top} - x_{bottom}}$$

Where $x_i$ is the living standard of country *i*, $x_{bottom}$ is the living standard of the poorest country in the region, and $x_{top}$ is the living standard of the richest country in the region.

Inequality between countries was quantified as the Gini coefficient, calculated as follows:

$$Gini\ coefficient = \frac{\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2\bar{x}}$$

Where $x_i$ and $x_j$ are the living standard of country *i* and *j*, $\bar{x}$ is the mean living standard in the region, and *n* is the number of countries in the region.

Advantageous and disadvantageous inequality between countries were calculated as follows:

$$Advantageous\ inequality = \frac{1}{n-1}\sum_{j=1}^{n} max|x_i - x_j, 0|$$

$$Disadvantageous\ inequality = \frac{1}{n-1}\sum_{j=1}^{n} max|x_j - x_i, 0|$$

Where $x_i$ and $x_j$ are the living standard of country *i* and *j*, and *n* is the number of countries in the region.

*Individual-level analysis.*

The individual-level analysis was conducted on individual responses from the Gallup World Poll. To assess the relationship between international comparisons and well-being we used regression analysis, controlling for

variables that have been shown in previous studies to affect well-being or could underlie the influence of international comparisons. We included three types of variables: (i) characteristics of an individual, (ii) characteristics of a country that individual lives in, (iii) characteristics of a region that the individual lives in. Additionally, the model included a fixed effect for the year of the survey, country and region of the world that allowed us to control for the effects specific for years or regions not accounted for by other variables. As country specific fixed effects could not be estimated for countries that did not change their rank within the observed time period, we limited our analysis to countries that moved up or down in their regional ranking between the year 2006 and 2018. Due to missing values for some of the variables for some countries or individuals, the individual-level analysis was based on a considerably smaller sample of 478669 responses. The supplementary material discusses alternative specifications of the above regression model, including: a) a sequential regression approach that decorrelates the absolute living standards and international comparisons, ensuring that the observed results are not stemming from a positive correlation between income and rank, or negative correlation between average income and inequality, b) an analysis including all countries and country-specific random-effects that relies on less conservative assumptions about the relationship between observed and unobserved variables (Plümper, & Troeger, 2007), but allowed us to maximize the number of data points used in the analysis.

Characteristics of an individual were based on the responses from the Gallup World Poll and included: log equivalized annual household income at PPP in international dollars (referred to household income), percentile that the given annual household income per capita belongs to within a country (referred to as income rank in the country), percentile that the given annual household income per capita belongs to in a region of the world (referred to as income rank in the region). Equivalization of household income was achieved by dividing total annual household income by the square root of the

number of people in the household. Equivalization was used to account for the fact that costs of living per person increase non-linearly with the household size, as recommended by OECD and used by other authors (Jebb et al., 2018; OECD, 2011). Percentile of income that the household belongs to was estimated based on the distribution of incomes in the Gallup Poll. To differentiate this measure from the income percentile in a person's country, income percentile in the region of the world was estimated based on all incomes except the incomes from the country that the person belongs to. In an example region that contains only two countries, regional income percentile represents what income percentile a person would belong to if they lived in the other country. As Gallup Poll contains around 1000 responses from each country, irrespective of the population of that country, for the estimation of income percentiles in the region we used weighted percentiles that assigned weights proportional to the population of a given country. This means that countries with more citizens had a bigger influence on the estimation of income percentile in the region than countries with lower number of citizens.

Additionally, we included gender, age (based on previous studies assumed to have non-linear quadratic effect), education (elementary, secondary, tertiary), marital status (single, domestic partner, married, divorced, widowed), employment status (full-time, self-employed, part-time, unemployed, out of workforce), religiousness (answer to question "Is religion an important part of your daily life?"), settlement size (rural, village, suburb of a city, city), health problems (answer to question "Do you have any health problems that prevent you from doing any of the things people your age normally can do?"), social support (answer to question "If you were in trouble, do you have relatives or friends you can count on to help you whenever you need them, or not?"), perceived freedom (answer to question "Are you satisfied or dissatisfied with Your freedom to choose what you do with your life?"), feeling safe (answer to question "Do you feel safe walking alone at night in the city or area where you live?"), perception of corruption (average response to

questions: "Is corruption widespread within businesses located in this country, or not?" and "Is corruption widespread throughout the government in this country, or not?"), and confidence in public institutions (average response to question "In (this country), do you have confidence in each of the following, or not?": military, the judicial system, national government, financial institutions, honesty of elections).

Characteristics of a country included total Gross Domestic Product at Purchasing Power Parity, Gross Domestic Product Growth, Log Gross National Income at Purchasing Power Parity per capita, unemployment rate, total natural resources rents (% of GDP), military expenses (% of GDP), Exports of goods and services (% of GDP), Imports of goods and services (% of GDP), number of refugees seeking asylum, number of refugees originating from a country (% of population), and net migration (linearly interpolated for missing years, due to the statistic being estimated only every 3 years) taken from World Bank national accounts database. Additionally, we included KOF Globalization Index that measures how interconnected a given country is with other countries on a social, economic and political level; Political Stability and Absence of Violence/Terrorism index (thereafter called Political stability index) from Worldwide Governance Indicators, measuring perceptions of the likelihood of political instability and/or politically-motivated violence; Voice and Accountability Index (thereafter called freedom index) from Worldwide Governance Indicators, that captures perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media; Henley Passport Index (thereafter called freedom of movement), that measures the number of countries citizens of a given country can enter without requiring a visa.

Characteristics of a region included averages in the region of Gross National Income at Purchasing Power parity per capita, Globalization Index, Political stability index and Freedom index. To account for possible effects of

other variables not included above, we also included fixed effects for each country, region and year.

To illustrate the effect of rank of a country and inequality of living standards within regions of the world, we plot the average residuals (difference between predicted and observed responses) for each analyzed year and country from a regression model including all the above control variables, and the effects of: (**Figure 4A**) absolute living standards and inequality between countries fixed to 0, and (**Figure 4B**) absolute living standards and rank of a country fixed to 0, where 0 is the mean value for standardized variables. To illustrate the effect of rank of a person in the region (**Figure 4C**), we plot the average residuals for each regional percentile of living standards, averaged separately for each country, from a regression model including all the above control variables, and the effects of personal living standard (log equivalized household income at PPP) and person's rank in the region set to 0, where 0 is the mean value for standardized variables.

*Relative importance analysis.*

To assess the magnitude of the observed effects, we perform relative importance analysis that allows quantifying the proportionate contribution each predictor makes to $R^2$, considering both its unique contribution and its contribution when combined with other variables (LeBreton and Tonidandel, 2008). It achieves this by transforming correlated predictors into new variables that are uncorrelated with each other but maximally correlated to their own respective original predictor variable. The results from this analysis overcome the limitations in the interpretability of standardized beta coefficients in regression models with multiple intercorrelated predictors.

## SUPPLEMENTARY INFORMATION

### Regional reference group assumption

To test the assumption that international comparisons are made within reference regions, rather than globally, we run a regression model predicting individual-level life satisfaction from the rank of a country, the rank of a person and inequality between countries without subdividing the world into regions, but including all control variables from the original model. For ease of comparison, this model restricts the analysis to countries included in the original model. The global international comparison model had a much worse fit (BIC = 2037799) than the model with world subdivisions (BIC = 2030512).

### Sensitivity analysis

To test the robustness of the observed effects of international comparisons, we run several different specifications of the regression model. Model (1) is the main model described in the manuscript. All other models are the same as Model (1), except the highlighted differences.

Models (2) and (5) address potential multicollinearity problem between main variables of interest and absolute income by decorrelating these variables prior to the analysis. To achieve that, we follow sequential regression approach (Dormann et al., 2012; Graham, 2003), in which all shared variance between variables is assigned to a variable considered more important. We construct the following hierarchy of variables, that gives primacy to absolute income:

*Log equivalized household income at PPP > Personal rank in the country > Personal rank in the region*

*Log Gross National Income per capita at PPP > Country's rank in the region*

*Average Gross National Income per capita at PPP > Inequality between the countries*

Variables higher in the hierarchy are assigned all the shared variance with variables lower in the hierarchy. This ensures that any observed effect of variables lower in the hierarchy cannot be explained by shared variance with variables higher in the hierarchy.

Model (3) includes all countries in the analysis, and instead of country-specific fixed effects, includes country-specific random-effects. This approach relies on less conservative assumptions about the relationship between observed and unobserved variables (Plümper, & Troeger, 2007), but allowed us to maximize the number of data points used in the analysis by including all countries. Models (4) and (5) include only control variables related to income distribution and income comparisons.

In all cases, we observe that the effect of personal rank in the country, personal rank in the region of the world, country's rank in the region of the world, and inequality between countries in the region of the world are significant, and of the same sign as in the main model.

**S1 Table. Regression models and robustness checks.** The table reports standardized beta coefficients and significance values from different specifications of the regression model. Model (1) is the main model reported in the manuscript, which includes all control variables, and fixed effects for the country, year and region. Model (2) and (4) takes a sequential regression approach, by decorrelating international comparison variables from variables describing income on the personal, country and regional level. Model (3) includes country-specific random effects instead of fixed-effects. Model (4) and (5) include only control variables related to income and income comparisons.

| **Models** | **(1)** | **(2)** | **(3)** | **(4)** | **(5)** |
|---|---|---|---|---|---|
| **Intercept** | 4.58* | 4.57* | 3.96* | 4.53* | 4.52* |
| **Personal rank in the country** | 0.21* | | 0.25* | 0.34* | |
| **Personal rank in the region** | 0.22* | | 0.18* | 0.21* | |
| **Country's rank in the region** | 0.18* | | 0.13* | 0.1* | |
| **Inequality between the countries** | -0.55* | | -0.46* | -0.63* | |

| | | | | | |
|---|---|---|---|---|---|
| **Personal rank in the country[a]** | | 0.37* | | | 0.47* |
| **Personal rank in the region[b]** | | 0.35* | | | 0.29* |
| **Country's rank in the region[c]** | | 0.18* | | | 0.1* |
| **Inequality between the countries[d]** | | -0.55* | | | -0.63* |
| | | | | | |
| **Log equivalized household income at PPP** | -0.07* | 0.15* | -0.06* | -0.09* | 0.19* |
| **Gender** | 0.14* | 0.14* | 0.14* | | |
| **Education level** | 0.18* | 0.18* | 0.17* | | |
| **Age** | -0.22* | -0.22* | -0.2* | | |
| **Age$^2$** | 0.1* | 0.1* | 0.11* | | |
| **Health problems** | -0.43* | -0.43* | -0.43* | | |
| **Number of children below age of 15** | 0.05* | 0.05* | 0.05* | | |
| **Unemployed** | -0.4* | -0.4* | -0.37* | | |
| **Self-employed** | -0.03* | -0.03* | -0.01 | | |
| **Retired** | 0.02* | 0.02* | 0.04* | | |
| **Employed part-time** | 0.01 | 0.01 | 0.03* | | |
| **Married** | 0.09* | 0.09* | 0.08* | | |
| **Divorced** | -0.17* | -0.17* | -0.2* | | |
| **Widowed** | -0.16* | -0.16* | -0.16* | | |
| **Domestic partner** | -0.11* | -0.11* | -0.12* | | |
| **Settlement size** | 0.05* | 0.05* | 0.05* | | |
| **Being religious** | 0.03* | 0.03* | 0.04* | | |
| **Social support** | 0.61* | 0.61* | 0.62* | | |
| **Perception of personal freedom** | 0.3* | 0.3* | 0.31* | | |
| **Perception of corruption** | -0.06* | -0.06* | -0.05* | | |
| **Confidence in public institutions** | 0.16* | 0.16* | 0.16* | | |
| **Feeling safe** | 0.15* | 0.15* | 0.15* | | |
| | | | | | |
| **Log Gross National Income per capita at PPP** | 0.25* | 0.31* | 0.43* | 0.03 | 0.06* |
| **Gini coefficient** | -0.1* | -0.1* | -0.18* | | |
| **Gross Domestic Product Growth** | 0 | 0 | -0.01 | | |
| **Inflation rate** | -0.03* | -0.03* | -0.04* | | |
| **Unemployment rate** | -0.18* | -0.18* | -0.16* | | |
| **Life expectancy** | 0.51* | 0.51* | 0.12* | | |
| **Imports** | -0.07* | -0.07* | 0.02 | | |
| **Exports** | 0.02 | 0.02 | 0 | | |
| **Value of natural resources** | -0.22* | -0.22* | -0.17* | | |

| | | | | | |
|---|---|---|---|---|---|
| **Globalisation Index** | -0.25* | -0.25* | -0.1* | | |
| **Net migration** | -0.15* | -0.15* | -0.11* | | |
| **Military expenses** | 0.02 | 0.02 | -0.01 | | |
| **Freedom of movement** | -0.08 | -0.08 | -0.08* | | |
| **Political stability index** | 0.05* | 0.05* | 0.08* | | |
| **Freedom index** | 0.17* | 0.17* | 0.1* | | |
| **Refugees seeking asylum** | 0.04* | 0.04* | 0.03* | | |
| **Refugees originating from country** | 0 | 0 | 0 | | |
| **Average Gross National Income per capita at PPP** | 0.92* | 1.19* | 0.7* | 0.07* | 0.38* |
| **Average political stability index** | 0.13* | 0.13* | 0.01 | | |
| **Average freedom index** | 0.06 | 0.06 | -0.22* | | |

All models additionally include an intercept, fixed-effects for each country, year and region of the world, with the exception of the model (4) that includes random-effects but not fixed effects for each country. Coefficients for country, year and region were omitted from the table. **a**: personal rank in the country decorrelated from log equivalized household income at PPP; **b**: personal rank in the region decorrelated from log equivalized household income at PPP and personal rank in the country; **c**: country's rank in the region decorrelated from log Gross National Income per capita at PPP, personal rank in the region and log equivalized household income at PPP; **d**: inequality between the countries decorrelated from average Gross National Income per capita at PPP in the region. * $p < 0.05$, **, $p < 0.01$, *** $p < 0.001$.

## Relative importance analysis

As it has been demonstrated that in the case of multiple correlated predictors multivariate regression approach often fails to correctly partition the variance to the correct predictors, estimates of the standardized beta coefficients might be misleading in informing which predictors are the most important. To address this problem, we conduct relative importance analysis, that has been developed to solve this issue (Johnson, 2000; LeBreton and Tonidandel, 2008). It quantifies the proportionate contribution each predictor makes to $R^2$, considering both its unique contribution and its contribution when combined with other variables. It achieves this by transforming correlated predictors into new variables that are uncorrelated with each other but maximally correlated to their own respective original predictor variable. The results from this analysis overcome the limitations in the interpretability

of standardized beta coefficients in regression models with multiple intercorrelated predictors. Table S2 reports the results from this analysis.

**S2 Table. Relative importance analysis.** The table reports the results from the relative importance analysis. The model explains in total 32.75% of the variance in individuals' life satisfaction. The column with the raw contribution to $R^2$ sums up to this value, when additionally including the contribution of fixed effects of country, year and region. The column with % contribution to $R^2$ reports what % of the total explained variance can be attributed to each predictor.

| | **Raw contribution to $R^2$** | **% Contribution to $R^2$** |
|---|---|---|
| **Social support (I)** | 0.0196 | 5.99 |
| **Personal rank in the region (I)** | 0.0191 | 5.84 |
| **Personal rank in the country (I)** | 0.0172 | 5.26 |
| **Education level (I)** | 0.0160 | 4.89 |
| **Health problems (I)** | 0.0121 | 3.69 |
| **Life expectancy (C)** | 0.0090 | 2.74 |
| **Log Gross National Income per capita at PPP** | 0.0085 | 2.59 |
| **Perception of personal freedom (I)** | 0.0081 | 2.46 |
| **Inequality between countries in the region (R)** | 0.0078 | 2.39 |
| **Freedom of movement (C)** | 0.0075 | 2.29 |
| **Globalization Index (C)** | 0.0073 | 2.24 |
| **Country's rank in the region (C)** | 0.0071 | 2.17 |
| **Freedom index (R)** | 0.0064 | 1.97 |
| **Average Gross National Income per capita at PPP in the region (R)** | 0.0064 | 1.96 |
| **Freedom index (C)** | 0.0062 | 1.88 |
| **Stability index (R)** | 0.0057 | 1.75 |
| **Confidence in public institutions (I)** | 0.0055 | 1.67 |
| **Log equivalized household income at PPP (I)** | 0.0050 | 1.54 |
| **Age (I)** | 0.0046 | 1.40 |
| **Settlement size (I)** | 0.0043 | 1.31 |
| **Inequality (C)** | 0.0040 | 1.23 |
| **Stability index (C)** | 0.0037 | 1.13 |
| **Value of natural resources (C)** | 0.0036 | 1.09 |
| **Perception of corruption (I)** | 0.0031 | 0.95 |
| **Unemployed (I)** | 0.0029 | 0.88 |
| **Being religious (I)** | 0.0025 | 0.77 |

| | | |
|---|---|---|
| **Feeling safe (I)** | 0.0022 | 0.66 |
| **Unemployment rate (C)** | 0.0021 | 0.63 |
| **Widowed (I)** | 0.0018 | 0.55 |
| **Exports (C)** | 0.0017 | 0.51 |
| **Net migration (C)** | 0.0016 | 0.50 |
| **Military expenses (C)** | 0.0014 | 0.43 |
| **Inflation (C)** | 0.0014 | 0.43 |
| **Refugees origination from country (C)** | 0.0012 | 0.37 |
| **Age$^2$ (I)** | 0.0009 | 0.28 |
| **Self-employed (I)** | 0.0009 | 0.27 |
| **Imports (C)** | 0.0009 | 0.27 |
| **Divorced (I)** | 0.0007 | 0.22 |
| **Refugees seeking asylum (C)** | 0.0004 | 0.14 |
| **Gender (I)** | 0.0004 | 0.13 |
| **GDP growth (C)** | 0.0004 | 0.12 |
| **Being married (I)** | 0.0004 | 0.12 |
| **Being retired (I)** | 0.0003 | 0.10 |
| **Number of children below the age of 15 (I)** | 0.0002 | 0.07 |
| **Working part-time (I)** | 0.0002 | 0.05 |
| **Having domestic partner (I)** | 0.0001 | 0.04 |

The table does not report fixed effects for the country, year and region. (I) individual-level variable, (C) country-level variable, (R) region-level variable.

# REFERENCES


Alesina, A., Di Tella, R., and MacCulloch, R. (2004). Inequality and happiness: Are Europeans and Americans different? *Journal of Public Economics*, 88(9), 2009–2042. https://doi.org/10.1016/j.jpubeco.2003.07.006

Al Ramiah, A., Hewstone, M., and Schmid, K. (2011). Social Identity and Intergroup Conflict. *Psychological Studies*, 56(1), 44–52. https://doi.org/10.1007/s12646-011-0075-0

Anand, S., & Sen, A. (1994). Sustainable Human Development: Concepts and Priorities (SSRN Scholarly Paper ID 2294664). Social Science Research Network. https://papers.ssrn.com/abstract=2294664

Becchetti, L., Castriota, S., Corrado, L., and Ricca, E. G. (2013). Beyond the Joneses: Inter-country income comparisons and happiness. *The Journal of Socio-Economics*, 45, 187–195. https://doi.org/10.1016/j.socec.2013.05.009

Bjørnskov, C., Dreher, A., Fischer, J. A. V., Schnellenbach, J., and Gehring, K. (2013). Inequality and happiness: When perceived social mobility and economic reality do not match. *Journal of Economic Behavior and Organization*, *91*, 75–92. https://doi.org/10.1016/j.jebo.2013.03.017

Boyce, C. J., Brown, G. D. A., and Moore, S. C. (2010). Money and happiness: Rank of income, not income, affects life satisfaction. *Psychological Science*, 21(4), 471–475. https://doi.org/10.1177/0956797610362671

Brown, G. D. A., Gardner, J., Oswald, A. J., and Qian, J. (2008). Does Wage Rank Affect Employees' Well-being? *Industrial Relations: A Journal of Economy and Society*, *47*(3), 355–389. https://doi.org/10.1111/j.1468-232X.2008.00525.x

Brown, S., Gray, D., and Roberts, J. (2015). The relative income hypothesis: A comparison of methods. *Economics Letters*, *130*, 47–50. https://doi.org/10.1016/j.econlet.2015.02.031

Campo, M., Mackie, D. M., and Sanchez, X. (2019). Emotions in Group Sports: A Narrative Review From a Social Identity Perspective. *Frontiers in Psychology*, 10. https://doi.org/10.3389/fpsyg.2019.00666

Capelli, C., and Vaggi, G. (2013). A better indicator of standards of living: The Gross National Disposable Income (No. 062). Retrieved from University of Pavia, Department of Economics and Management website: https://ideas.repec.org/p/pav/demwpp/demwp0062.html

Clark, A. E., & D'Ambrosio, C. (2015). Chapter 13 - Attitudes to Income Inequality: Experimental and Survey Evidence. In A. B. Atkinson & F. Bourguignon (Eds.), Handbook of Income Distribution (Vol. 2, pp. 1147–1208). Elsevier. https://doi.org/10.1016/B978-0-444-59428-0.00014-X

Clark, A. E., and Senik, C. (2010). Who compares to whom? The anatomy of income comparisons in Europe. *The Economic Journal*, *120*(544), 573–594. Retrieved from JSTOR.

De Neve, J.-E., Ward, G., De Keulenaer, F., Van Landeghem, B., Kavetsos, G., and Norton, M. I. (2018). The Asymmetric Experience of Positive and Negative Economic Growth: Global Evidence Using Subjective Well-Being Data. *The Review of Economics and Statistics*, 100(2), 362–375. https://doi.org/10.1162/REST_a_00697

Delhey, J., and Kohler, U. (2006). From Nationally Bounded to Pan-European Inequalities? On the Importance of Foreign Countries as Reference Groups. *European Sociological Review*, 22(2), 125–140. https://doi.org/10.1093/esr/jci047

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., McClean, C., Osborne, P. E., Reineking, B., Schröder, B., Skidmore, A. K., Zurell, D., and Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Eccleston, C. P., & Major, B. N. (2016). Attributions to Discrimination and Self-Esteem: The Role of Group Identification and Appraisals: Group Processes & Intergroup Relations. https://doi.org/10.1177/1368430206062074

Ellemers, N., van Knippenberg, A., and Wilke, H. (1990). The influence of permeability of group boundaries and stability of group status on strategies of individual mobility and social change. *The British Journal of Social Psychology*, *29 ( Pt 3)*, 233–246. https://doi.org/10.1111/j.2044-8309.1990.tb00902.x

Ellemers, N., Spears, R., and Doosje, E. J. (1997). Sticking together or falling apart: Ingroup identification as a psychological determinant of group commitment versus individual mobility. *Journal of Personality and Social Psychology*, 72. https://doi.org/https://doi.org/10.1037/0022-3514.72.3.617

Fehr, E., and Schmidt, K. M. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, *114*(3), 817–868. Retrieved from JSTOR.

Ferrer-i-Carbonell, A. (2005). Income and well-being: An empirical analysis of the comparison income effect. Journal of Public Economics, 89(5), 997–1019. https://doi.org/10.1016/j.jpubeco.2004.06.003

Firebaugh, G., and Schroeder, M. B. (2009). Does your neighbor's income affect your happiness? *AJS; American Journal of Sociology*, *115*(3), 805–831. https://doi.org/10.1086/603534

Graham, M. H. (2003). Confronting Multicollinearity in Ecological Multiple Regression. *Ecology*, 84(11), 2809–2815. https://doi.org/10.1890/02-3114

Hagerty, M. R. (2000). Social comparisons of income in one's community: evidence from national surveys of income and happiness. *Journal of Personality and Social Psychology*, *78*(4), 764–771. https://doi.org/10.1037//0022-3514.78.4.764

Heliwell, J. F., Huang, H., and Wang, S. (2019). Changing world happiness. In: J. F. Heliwell, R. Layard and J.D. Sachs (Eds.), World happiness report. New York: Sustainable Development Solutions Network.

Helson, H. (1964). *Adaptation-level theory* (Vol. xvii). Oxford, England: Harper & Row.

Hirschman, A. O., and Rothschild, M. (1973). The Changing Tolerance for Income Inequality in the Course of Economic Development. *The Quarterly Journal of Economics*, *87*(4), 544–566. https://doi.org/10.2307/1882024

Jebb, A. T., Tay, L., Diener, E., and Oishi, S. (2018). Happiness, income satiation and turning points around the world. *Nature Human Behaviour*, 2(1), 33–38. https://doi.org/10.1038/s41562-017-0277-0

Johnson, J. W. (2000). A Heuristic Method for Estimating the Relative Weight of Predictor Variables in Multiple Regression. Multivariate Behavioral Research, 35(1), 1–19. https://doi.org/10.1207/S15327906MBR3501_1

Kuziemko, I., Buell, R. W., Reich, T., & Norton, M. I. (2014). "Last-Place Aversion": Evidence and Redistributive Implications. The Quarterly Journal of Economics, 129(1), 105–149. https://doi.org/10.1093/qje/qjt035

Lahusen, C., & Kiess, J. (2019). 'Subjective Europeanization': Do inner-European comparisons affect life satisfaction? European Societies, 21(2), 214–236. https://doi.org/10.1080/14616696.2018.1438638

LeBreton, J. M., and Tonidandel, S. (2008). Multivariate relative importance: extending relative weight analysis to multivariate criterion spaces. *The Journal of Applied Psychology*, *93*(2), 329–345. https://doi.org/10.1037/0021-9010.93.2.329

Luhtanen, R., and Crocker, J. (1992). A Collective Self-Esteem Scale: Self-Evaluation of One's Social Identity. *Personality and Social Psychology Bulletin*, 18(3), 302–318. https://doi.org/10.1177/0146167292183006

Luttmer, E. F. P. (2005). Neighbors as Negatives: Relative Earnings and Well-Being. *The Quarterly Journal of Economics*, *120*(3), 963–1002. https://doi.org/10.1093/qje/120.3.963

Milanovic, B. (2014). Global Inequality of Opportunity: How Much of Our Income Is Determined by Where We Live? *The Review of Economics and Statistics*, *97*(2), 452–460. https://doi.org/10.1162/REST_a_00432

Mummendey, A., Kessler, T., Klink, A., and Mielke, R. (1999). Strategies to cope with negative social identity: predictions by social identity theory and relative deprivation theory. *Journal of Personality and Social Psychology*, *76*(2), 229–245. https://doi.org/10.1037//0022-3514.76.2.229

Mussweiler, T., Gabriel, S., and Bodenhausen, G. V. (2000). Shifting social identities as a strategy for deflecting threatening social comparisons. *Journal of Personality and Social Psychology*, 79(3), 398–409. https://doi.org/10.1037//0022-3514.79.3.398

Oishi, S., Kesebir, S., and Diener, E. (2011). Income Inequality and Happiness. *Psychological Science*, *22*(9), 1095–1100. https://doi.org/10.1177/0956797611417262

Organization for Economic Co-operation and Development. (2011). Divided we stand: Why inequality keeps rising. Paris: OECD.

Parducci, A. (1995). Happiness, pleasure, and judgment: The contextual theory and its applications. Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Pérez-Asenjo, E. (2011). If happiness is relative, against whom do we compare ourselves? Implications for labour supply. *Journal of Population Economics*, *24*(4), 1411–1442. Retrieved from JSTOR.

Plümper, T., & Troeger, V. E. (2007). Efficient Estimation of Time-Invariant and Rarely Changing Variables in Finite Sample Panel Analyses with Unit Fixed Effects. Political Analysis, 15(2), 124–139. https://doi.org/10.1093/pan/mpm002

Powdthavee, N., Burkhauser, R. V., and De Neve, J.-E. (2017). Top incomes and human well-being: Evidence from the Gallup World Poll. *Journal of Economic Psychology*, *62*, 246–257. https://doi.org/10.1016/j.joep.2017.07.006

Rangel, A., and Clithero, J. A. (2012). Value normalization in decision making: theory and evidence. *Current Opinion in Neurobiology*, 22(6), 970–981. https://doi.org/10.1016/j.conb.2012.07.011

Rutledge, R. B., de Berker, A. O., Espenhahn, S., Dayan, P., and Dolan, R. J. (2016). The social contingency of momentary subjective well-being. *Nature Communications*, *7*(1), 1–8. https://doi.org/10.1038/ncomms11825

Sági, M. (2011). Determinants of Satisfaction with Living Standards in Transition Societies. *International Journal of Sociology*, 41(4), 55–78. https://doi.org/10.2753/IJS0020-7659410403

Salice, A., and Montes Sánchez, A. (2016). Pride, Shame, and Group Identification. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00557

Sen, S. A. and A. (1994). *Human development Index: Methodology and Measurement* (No. HDOCPA-1994-02). Retrieved from Human Development Report Office (HDRO), United Nations Development Programme (UNDP) website: https://ideas.repec.org/p/hdr/hdocpa/hdocpa-1994-02.html

Siem, B., Oettingen, M. von, Mummendey, A., and Nadler, A. (2013). When status differences are illegitimate, groups' needs diverge: Testing the needs-based model of reconciliation in contexts of status inequality. *European Journal of Social Psychology*, 43(2), 137–148. https://doi.org/10.1002/ejsp.1929

Smith, H. J., and Tyler, T. R. (1997). Choosing the Right Pond: The Impact of Group Membership on Self-Esteem and Group-Oriented Behavior. *Journal of Experimental Social Psychology*, *33*(2), 146–170. https://doi.org/10.1006/jesp.1996.1318

Soltani, A., Martino, B. D., and Camerer, C. (2012). A Range-Normalization Model of Context-Dependent Choice: A New Model and Evidence. *PLOS Computational Biology*, *8*(7), e1002607. https://doi.org/10.1371/journal.pcbi.1002607

Stewart, N., Chater, N., & Brown, G. D. A. (2006). Decision by sampling. Cognitive Psychology, 53(1), 1–26. https://doi.org/10.1016/j.cogpsych.2005.10.003

Osborne, D., Sibley, C. G., Huo, Y. J., and Smith, H. (2015). Doubling-down on deprivation: Using latent profile analysis to evaluate an age-old assumption in relative deprivation theory. *European Journal of Social Psychology*, 45(4), 482–495. https://doi.org/10.1002/ejsp.2099

Wasserman, L. (2000). Bayesian Model Selection and Model Averaging. *Journal of Mathematical Psychology*, *44*(1), 92–107. https://doi.org/10.1006/jmps.1999.1278

Whelan, C. T., and Maître, B. (2009a). Europeanization of inequality and European reference groups. *Journal of European Social Policy*, 19(2), 117–130. https://doi.org/10.1177/0958928708101865

WHELAN, P. C. T., AND MAÎTRE, B. (2009B). The 'europeanisation' of references groups. European societies, 11(2), 283–309. HTTPS://DOI.ORG/10.1080/14616690701846938

# Chapter 5

# When We Want To Know What We Already Know: Computational Mechanisms Underlying Self-verifying Information-seeking.

**Filip Gesiarz[*1], Tali Sharot[*1]**

**1**Affective Brain Lab, Department of Experimental Psychology, University College London, London, UK

* to whom correspondence should be addressed: filip.gesiarz.15@ucl.ac.uk, t.sharot@ucl.ac.uk

## ABSTRACT

Many studies have demonstrated that we often prefer self-relevant information that confirms our beliefs over information that contradicts them – a motive known as self-verification. However, it is still unclear how do people decide what source of information will confirm their beliefs and what drives them to seek such confirmation. To shed new light on these questions, we quantified the surprise signal accompanying information-seeking in a new experimental task, in which people repeatedly face an opportunity to reveal either positive or negative information about themselves. We find that experiencing a surprise when revealing information about oneself induces both negative hedonic and cognitive reactions, even when the information is positive. Individual differences between people in these reactions suggest that self-verification behavior is primarily driven by a decrease in self-evaluation confidence in response to surprise. We uncover a variety of heuristics and reinforcement-learning mechanisms that people use while looking for information that minimizes such surprises and find evidence implying that they select strategies that are the most optimal for their general self-esteem and specific circumstances. A reference-point heuristic based on average self-evaluations stands out as the most important strategy employed by participants. Our findings provide a first step in creating a computational theory of self-verifying behavior.

## INTRODUCTION

With the current abundance of information, it has never been easier to be selective about the content that one would like to be exposed to. Previous studies have demonstrated that people often prefer information that confirms their core beliefs, rather than information that challenges them (Pettit and Joiner, 2001; Kwang and Swann, 2010; Kappes et al., 2020). Despite common demonstrations of this behavior, it is still unclear what drives these preferences, and how do people decide what source of information would confirm their beliefs when the content of information is uncertain. To answer these questions, we used reinforcement learning framework (Sutton and Barto, 2018), that assumes that agents update their beliefs by learning from a discrepancy between their predictions and observed outcomes (i.e. prediction errors), and free-energy principle (Friston, 2010) according to which agents are motivated in their actions to minimize this discrepancy in absolute terms (i.e. surprises). We investigate how prediction errors and surprise signals described in these frameworks map to hedonic and cognitive responses to (dis)confirmatory information about oneself. Furthermore, we test which possible heuristics or learning mechanisms that maximize positive prediction errors or minimize surprises can explain how people choose sources of information when its content is uncertain. We describe a new implementation of reference-point heuristic (Tversky and Kahneman, 1992) based on average evaluations of self and others, and show that people apply it in their information seeking choices when attempting to confirm their beliefs. Our findings provide a bridge between descriptive theories of information seeking developed in social psychology and a more mechanistic approach represented in decision neuroscience.

There are two dominant theories in social psychology that describe what self-relevant information people will be interested in knowing: self-enhancement theory (for review: Blaine and Crocker, 1993; Leary, 2007) and self-verification theory (Swann, 1983; for review: Leary, 2007). According to the

former, people will be motivated to hold the best possible view of themselves and therefore should always look for positive information. According to the latter, people are motivated to hold a coherent view of themselves, and therefore will look for information that confirms their core beliefs. Predictions of these theories will converge in the case of people with positive self-views but will diverge for people with-negative self-views, for whom the self-enhancement theory would predict a preference for positive information, while self-verification theory would predict a preference for negative information. A meta-analysis of studies dedicated to test these theories concluded that there is strong evidence for both (Kwang and Swann, 2010), however self-verification predominantly affects how people seek and evaluate the provided information, while self-enhancement affects how people feel about the provided information. Following these findings, we focus on two aspects of reactions to information: hedonic utility, measured by self-ratings of momentary feelings, and cognitive utility, measured by updates of confidence in self-views.

Most of the previous studies have focused on general self-esteem rather than specific self-views about oneself (for review: Kwang and Swann, 2010; but see: Swann, Pelham and Krull, 1989; Dutton, Brown and Jonathon, 1997; Chen, English, and Peng, 2006; Bernichon, Cook and Brown, 2003) despite dissociable effects of the two (Rosenber et al., 1995), and to the best of our knowledge, none attempted to precisely quantify the value that people aim to maximize/minimize while choosing self-verifying information. The present study fills in this gap, by developing a task in which participants repeatedly face an opportunity to confirm or enhance their specific view about themselves, by revealing either higher or lower ratings received from other participants about their personality traits. We suggest that self-verification is partly driven by a mechanism that aims to minimize experienced surprise when being exposed to information. Based on this assumption, we compute the magnitude of successes and failures of self-verification as an absolute

difference between received information and expected information based on beliefs about oneself. This allows us to assess if such computations are reflected in cognitive and hedonic responses to received information, as opposed to just looking at differences in responses to positive and negative feedback of low and high self-esteem groups, as in previous studies.

The idea that the human brain aims to minimize surprises has been suggested before in the free energy principle (Friston, 2010), according to which actions that minimize surprises help agents to build better representations of the world. This principle primarily applies to neural processes, but there are undergoing efforts to link it to a broad array of decision-making behaviors (Friston et al., 2013; Friston, 2018), and similarity between predictions of self-verification and free-energy principle has been explicitly noted before (Friston, 2018). Based on this principle, we test few possible implementations of policies minimizing surprises: one based on the learned information about the environment and the other based on learned consequences of actions (Friston et al., 2016; Daw et al., 2011; Daw, Niv and Dayan, 2005; Gläscher et al., 2010; Gesiarz and Crockett, 2015). We compare these learning mechanism to reliance on decision-making heuristics (Tversky and Kahneman, 1992).

To model participants learning behaviour, we use two variations of the Rescorla-Wagner algorithm: one that shares features with the model-based learning by simulating surprise following one's actions given the beliefs about the average feedback and self-evaluation on a given trait, and other that shares features with model-free learning by just learning the average experienced surprise associated with different actions in the past (Friston et al., 2016; Daw et al., 2011; Daw, Niv and Dayan, 2005; Gläscher et al., 2010; Gesiarz and Crockett, 2015). In model-based learning agents learn about the structure of their environment, simulate the consequences of each possible action, and choose the action with the highest simulated value. In the context of surprise minimization, model-based learning would aim to choose information source

which will result in the lowest expected surprise, based on what the agent knows about the environment. Model-free learning relies on comparisons of the accumulated value of outcomes experienced after taking each possible action, without any consideration of the environment. In the context of surprise minimization, model-free learning would aim to avoid information sources that lead to surprises in the past, and seek information sources that were associated with low surprises.

Another possibility is that people use simple rules that minimize surprises in most situations, without the need for learning (Gigerenzer and Todd, 1999). One example could be reliance on reference points that are often defined as an average expected value - a heuristic used in many areas of decision-making (Tversky and Kahneman, 1992; Wang and Johnson, 2012; Koop and Johnson, 2011). If a person is motivated to seek information that minimizes the risk of being surprised, they could rely on their average self-evaluation as a reference point, and seek positive information whenever their specific self-view is better than this reference point, or seek negative information whenever their specific self-view is worse than this reference point. An alternative could be that we treat how other people are evaluated on average as a reference point, recognizing that they private self-evaluations might be different from how people are perceived in general (Baumeister and Hutton, 1987; Tesser and Paulhus, 1983), and assuming that our public perception does not differ from a perception of an average person.

## RESULTS

We invited participants in groups of three and asked them to get to know each other by having a 10 minutes conversation with each person about three provided topics that aimed to induce broader familiarity with the other person. Subsequently, we asked participants to rate themselves and others on a list of 50 adjectives, on a scale ranging from 0 to 10, and rate how confident were they in their self-evaluations. In the main part of the experiment, during

50 trials corresponding to 50 different adjectives, participants had an opportunity to reveal either the higher rating (positive information) or the lower rating (negative information) that they received from other people. It was not possible to reveal both, and the identity of the person from which the rating came from was not revealed to the participant. Finally, after receiving information about how they were rated by other people, we asked participants again to rate themselves and provide confidence ratings in their self-evaluations on the same adjectives.

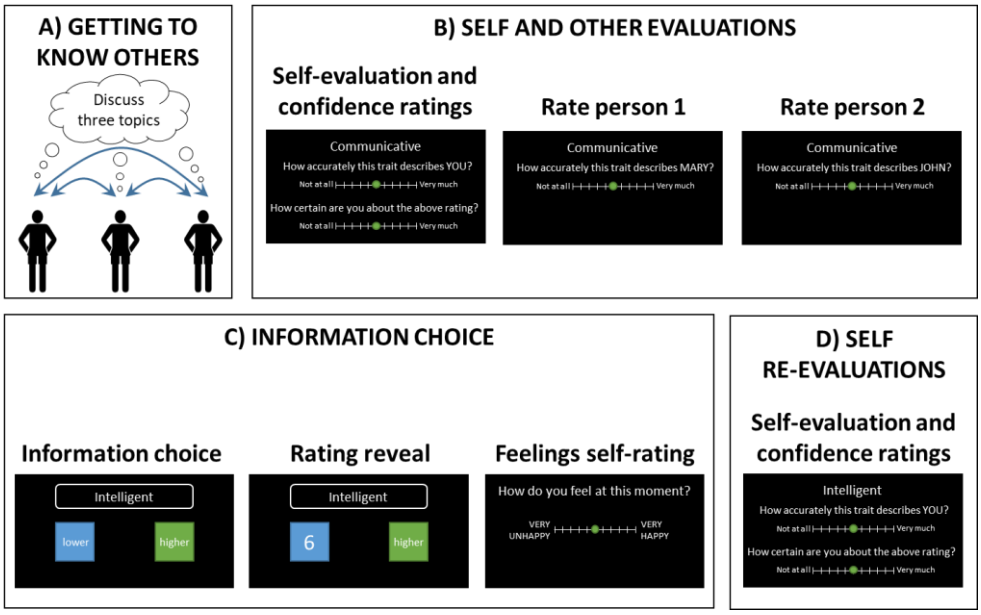


**Figure 1. Behavioural task.** The task consisted of four parts. (A) In the first part participants in a group of three were introduced to each other and were asked to talk in pairs for 10 minutes about three provided topics that aimed to induce a broader familiarity with another person. (B) Subsequently participants retired to separate cubicles, where they were asked to rate themselves and two other people on a list of 50 adjectives. Additionally, they were asked to provide a confidence rating in their self-evaluation. (C) In the main part of the experiment, participants had an opportunity to reveal how they were rated by others. They could either reveal the higher or the lower rating that they received on a specific trait. After they were informed about their rating, they were asked about how they feel at the current moment. (D) In the last part, they were asked again to provide ratings of themselves and their level of confidence in their rating. Parts (B), (C) and (D) each consisted of 50 trials, corresponding to 50 different adjectives, and started after the previous part was completed.

**People feel happier and more confident in their self-evaluations after receiving unsurprising information vs. surprising information**

First, we wanted to investigate the impact of receiving surprising feedback. To test if surprise influences the participant's momentary feelings and confidence in self-evaluations, we quantified surprise as an absolute difference between specific self-ratings and the received information. We used Generalized Linear Mixed Effects Model (GLME) to predict fluctuations in momentary feelings and changes in confidence in self-ratings based on experienced surprise with the received information, while controlling for the signed prediction error (signed difference between received information and self-ratings) and the value of received information. We find that the more surprising the information the less happy a person is ($\beta = -0.02$, $p < 0.001$) and the less confident they are in their subsequent self-evaluations ($\beta = -0.24$, $p < 0.001$). Additionally, the more positive the prediction error ($\beta = 0.02$, $p < 0.01$) and the more positive received information ($\beta = 0.05$, $p < 0.001$) the greater momentary happiness. However, prediction errors and information valence were not significantly related to changes in confidence (prediction errors: $\beta = 0.08$, $p = 0.18$; information value: $\beta = -0.05$, $p = 0.075$). These findings suggest that surprise has both negative hedonic and negative cognitive value, which possibly motivate people to minimize it (**Figure 2**).

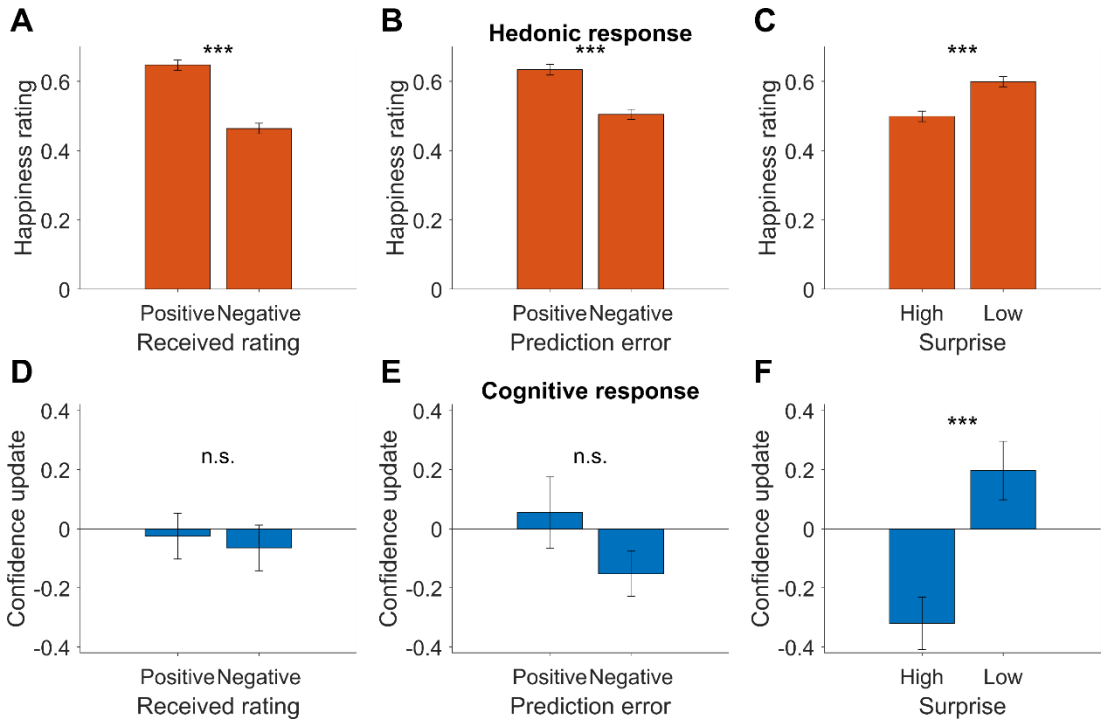


**Figure 2. Hedonic and cognitive responses to ratings, prediction errors and surprises.** To illustrate the effect of different aspects of received information on hedonic and cognitive responses, we dichotomized continuous variables based on median split in the case of received rating and surprise and based on the sign in the case of prediction error. Positive ratings (A) and prediction errors (B) were associated with higher happiness ratings, but were not associated with confidence updates (D), (E). High surprise was associated with lower happiness ratings and decreased in confidence. The statistics were computed using continuous variables using GLME. *** $p < 0.001$, n.s., not significant. Error bars represent SEM.

A previous study has suggested that receiving information of valence inconsistent with one's general self-esteem can reduce self-clarity, defined as stability, internal consistency and confidence in one's general self-concept (Stinson, Holmes, and Forest, 2010). To test if deviations between revealed information and general self-esteem were also reflected in hedonic and cognitive responses measured in our task, we run a GLME model with the following predictors: (i) general surprise, defined as an absolute difference between revealed rating and general self-esteem (average self-rating), (ii) general prediction error, defined as signed difference between received

information and general self-esteem (average self-rating), and (iii) the value of received information. General surprise and general prediction errors did not significantly predict hedonic (general surprise: $\beta$ = -0.00, p = 0.71; general prediction error: $\beta$ = 0.01, p = 0.86) nor the cognitive responses (general surprise: $\beta$ = 0.02, p = 0.58; general prediction error: $\beta$ = 0.18, p = 0.12), in contrast to surprises based on specific self-views as described above. This suggests that the potential effect of received ratings on the concept of self-clarity cannot explain the observed effects of surprises about the received ratings vs. specific self-views.

**Confirmatory information-seeking is primarily driven by heuristics**

So far we focused on the consequences of receiving information. Next, we investigate how people seek information. We find that participants were as likely to choose positive as a negative information source, suggesting that self-enhancement motive did not dominate the self-verification motive in our task ($t(61) = -0.13$, $p = 0.90$). According to the self-verification theory, people with high self-esteem should seek positive information, and people with low self-esteem should seek negative information (Swann, 1983). Consistently with this suggestion, we find that the average proportion of positive to negative information choices significantly correlates with general self-esteem ($R = 0.44$, $p < 0.001$). Predictions of the self-verification theory are much less clear with regard to specific self-views. Assuming that people are motivated to minimize surprises, those with generally low self-esteem should sometimes seek positive information, and those with generally high self-esteem should sometimes seek negative information, depending on what they expect to minimize surprise. We considered two groups of choice mechanisms that people might use to achieve this goal: (i) heuristics and (ii) reinforcement learning processes.

Among heuristics, we tested three possibilities. First, people could simply use the absolute scale, and seek positive information about qualities that they rate themselves high and seek negative information about qualities

that they rate themselves low. An alternative could assume that people have a reference point (Tversky and Kahneman, 1992), that defines what level of a trait is positive and what level of it is negative. We test two possibilities of such reference points: average self-evaluation, and average evaluation of others. Among learning mechanisms, we consider two strategies that people might employ described previously in the reinforcement literature (Friston et al., 2016; Daw et al., 2011): model-based learning, that simulates consequences of actions based on the learned structure of the environment, and model-free learning, that relies on learned values of different actions. We also consider that people might use a combination of both learning mechanisms – as often observed in different learning tasks (Daw et al., 2011; Daw, Niv and Dayan, 2005). Additionally, we tested versions of the above learning mechanisms that aim to maximize positive prediction errors, rather than surprises. To compare these different models, we used the Bayesian Information Criterion that simultaneously assesses model fit and parsimony (Smith and Spiegelhalter, 1980). The results suggest that heuristics in general fit closer to participants' behavior than reinforcement learning algorithms, and learning mechanisms that aim to minimize surprise fit better to the data that mechanisms that aim to maximize positive prediction errors.

Out of all heuristics, a heuristic that relied on reference-point based on average self-evaluation outperformed all other alternatives (**Figure 4D**). According to this heuristic, people are more likely to seek positive information about traits that they evaluate as higher than their average self-evaluation, and more likely to seek negative information about traits that they evaluate as lower than their average self-evaluation (**Figure 4F**). Importantly, we observed a substantial heterogeneity between participants. Although overall reference-point heuristic based on average self-evaluations was the best fitting strategy, when considering all participants together, reference-point heuristic based on average evaluations of others was identified as the most frequent best-fitting model when choosing the best model characterizing each individual (29.2% vs.

26.2%; see **Figure 4E**). Furthermore, for 44.6% of participants a learning process was identified as the best fitting strategy. For none of the participants choosing a source of information just based on one's self-evaluations was a better model than all other alternatives.

To investigate which strategy would be the most optimal in our task, we performed a simulation of choice behavior that aims to minimize surprise in a similar environment to the one used in the experiment (**Figure 4A – C**). This exercise shows that reference-point heuristic is an optimal strategy when the average self-esteem is close to an average evaluation received from others, outperforming model-based and model-free learning algorithms. However, when general self-esteem deviates substantially from the average rating received from the two sources of information, learning mechanisms will outperform reference-based heuristic. This pattern is consistent across different specifications of the environment and learning algorithms (see supplementary material). To test if such deviations might have prompted participants to employ learning processes over heuristics, we compared participants who primarily used heuristics with participants who primarily used learning processes on a measure of absolute distance from mean expected rating from the two sources of information.

Consistent with the above simulation, we find that participants who employed learning processes had a significantly higher absolute deviation from the mean expected rating than participants who relied on heuristics ($t(63) = 2.19$, $p = 0.03$). Participants who used learning strategies tended to change their preferred source of information in comparison to last trial after receiving surprising feedback ($t(26) = -1.78$, $p = 0.086$; **Fig 3.**), but participants using heuristics did not ($t(30) = -0.17$, $p = 0.87$; **Fig 3.**).

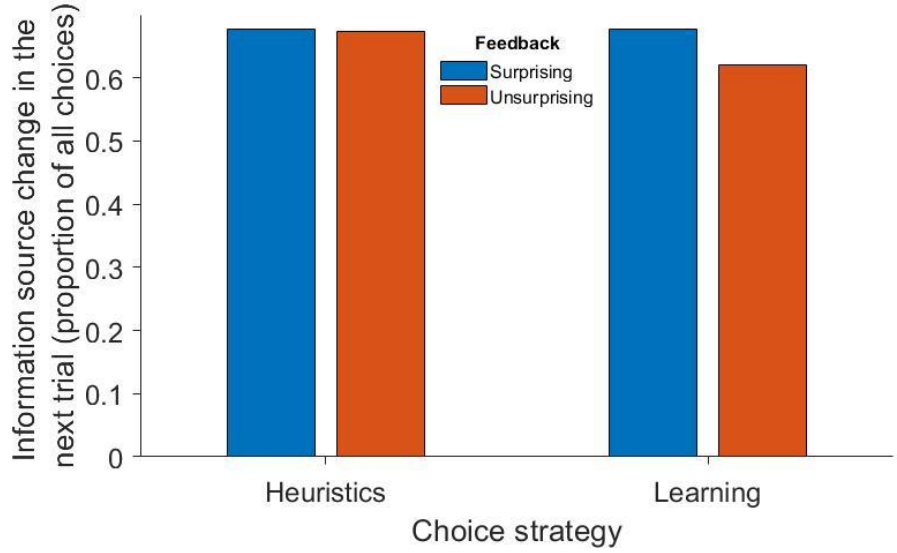


**Figure 3. Proportion of changes in preferred source of information (higher rating or lower rating) from trial to trial after receiving surprising or unsurprising feedback in the subgroup of participants who used heuristics in their choices, or learning mechanisms.**

This suggests that participants tend to use strategies that are optimal in minimizing surprises for their level of general self-esteem and expected mean information in the environment. Importantly, these two groups did not differ with respect to other characteristics, including general self-esteem ($t(63)= 0.81$, $p = 0.42$), proportion of positive information choices ($t(63)= 1.76$, $p = 0.08$), choice variability ($t(63)= 0.99$, $p = 0.33$), or experienced surprises ($t(63)= 0.99$, $p = 0.33$), limiting other possible explanations of the observed differences in employed strategies between participants.

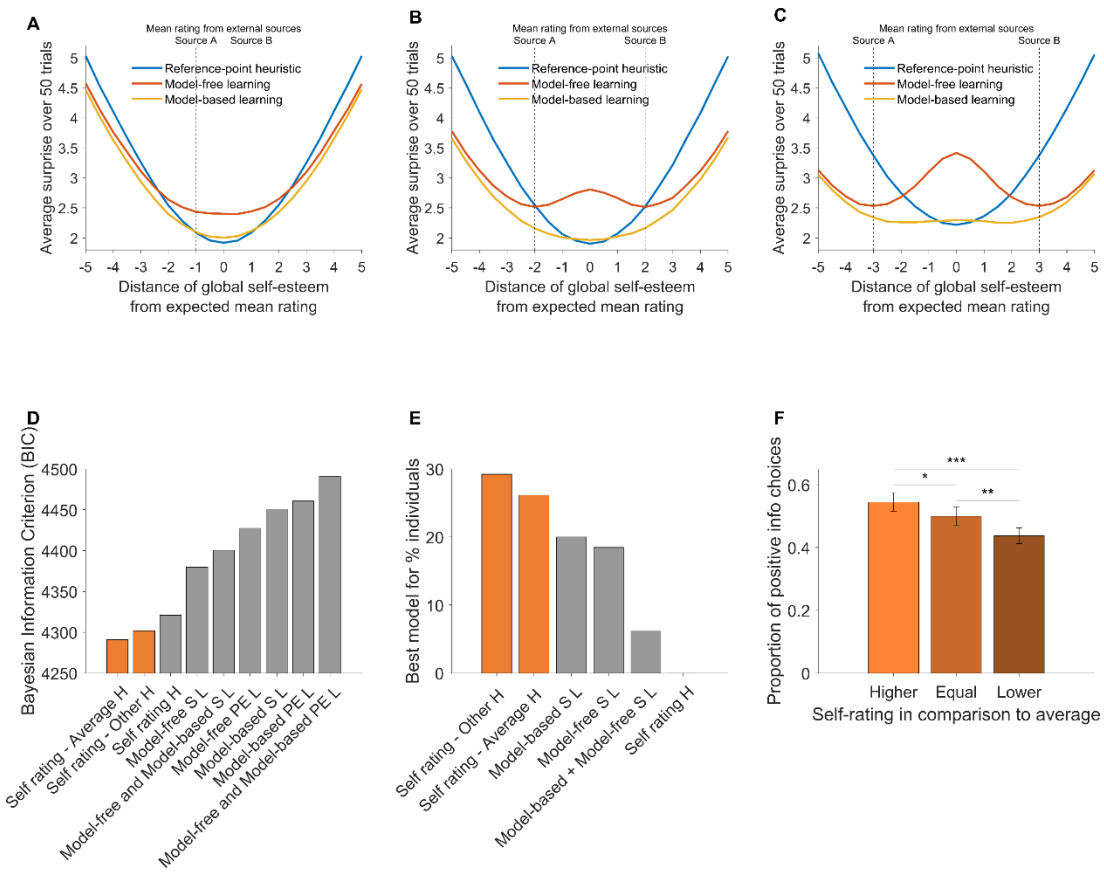


**Figure 4. Choice heuristics and learning mechanisms in simulations and behavior.** To investigate the optimality of different choice strategies that aim to minimize surprises, we created simulated agents that used either a reference-point heuristic, model-based learning or model free-learning in their choices, in an environment similar to the one used in our task. The agents could be characterized by different level of general-self esteem, ranging from deviating -5 from the average expected rating from the two sources of information to +5. Their specific self-evaluations on 50 traits were normally distributed around this general self-esteem. Each agent started with no knowledge of the environment, except a knowledge that one source of information is more positive than the other. We created 2000 agents per each algorithm and general self-esteem level. We then averaged the experienced surprise during the task over 50 trials, and averaged over agents separately for each algorithm and general self-esteem level. Plots (A), (B), and (C) demonstrate the same simulation, with different discrepancy between mean rating expected from each source of information (represented by a dashed line). This exercise shows that the reference-based heuristic leads to the lowest experienced surprise, whenever general self-esteem is close to the expected average rating from the two sources of information. Strategies that are based on learning outperform the heuristic the more the general self-esteem deviated from this average and the more the two sources of information disagree with each other. Plot (D) shows the summed Bayesian Information Criterion of models fitted to behavior of all participants. It indicates that the

reference-point heuristics (in orange) outperformed all other alternatives. Plot (E) shows percentage of participants for whom each of the models was the best model. Plot (F) illustrates the effect of reference-point heuristic based on average self-evaluations. It shows the proportion of positive information choices, whenever specific rating was higher, similar (general self-esteem − 0.5 < specific self-rating < general self-esteem + 0.5), or lower than general self-esteem. *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; H, heuristic; L, learning; S, surprise; PE, prediction error. Error bars represent SEM.

Next, we tested how the winning heuristic model relates to reaction times. Importantly, we did not find significant differences in reaction times between positive vs. negative information choices ($\beta$ = -0.01, $p$ = 0.62), nor between confirmatory vs. disconfirmatory information choices, as defined by the reference-point heuristic ($\beta$ = -0.01, $p$ = 0.72). Therefore, we wanted to investigate if reaction times are better characterized by quantities used in the reference-point heuristic, that is the distance of a self-rating from the mean self-rating. One possibility is that the probability of revealing positive information predicted by the model could be linearly related to logarithm of reaction time, similarly to other studies that demonstrated that selecting stimuli associated with positive affect is characterized by faster reaction times than selecting stimuli associated with negative affect (Leppänen, Tenhunen and Hietanen, 2003; Guitart-Masip, et al., 2011; Gesiarz, Cahill, and Sharot, 2019). Another possibility is that the logarithm of reaction time could be inversely related to decision uncertainty, which peaks at 50% choice probability, similarly to other studies showing that choice difficulty prolongs the decision time (Hong and Beck, 2010; Yu and Dayan, 2005; Gesiarz, Cahill, and Sharot, 2019). To test these possibilities, we fitted the winning reference-point heuristic model to all participants and extracted the estimated probability of choosing a positive information source for each trial and participant. We then used a GLME model to predict logarithm of reaction time from the probability of choosing positive information and choice uncertainty. We find that the probability of choosing positive information estimated based

on the heuristic model is inversely related to reaction time ($\beta$ = -0.06, $p < 0.001$). That is, the higher the self-rating than the average self-rating, the faster the choice, suggesting that despite lack of bias in the proportion of positive to negative choices, people might still exhibit a positivity bias in reaction times. Decision uncertainty did not significantly predict reaction times ($\beta$ = 0.01, $p = 0.34$). Additionally, we find that the more a participant relied on the reference-point heuristic (as estimated by the weight put on the heuristic rule in the choice model, when fitted individually to all participants), the faster the made their decisions on average ($R$ = -0.25, $p = 0.04$), consistent with the idea that heuristics are automatic in nature. We note that we did not find a difference in reaction times between a groups of participants using learning strategies and heuristics identified in previous paragraph ($t(63) = 1.56$, $p = 0.12$), suggesting that a weight measure that characterizes reliance on reference-point heuristic on a continuous rather than categorical scale might be required to capture the differences in reaction times.

How this finding relates to the prediction of the self-verification theory that people with high and low general self-esteem should seek positive and negative information respectively? A generalized mixed-effects model that predicted choices from general self-esteem and the reference-point heuristic reveals that these mechanisms are orthogonal, with no interaction between them, and the effect of general self-esteem being significant only at a trend level once controlled for the heuristic usage (general self-esteem: $\beta$ = 0.20, $p = 0.08$; reference-point heuristic: $\beta$ = 0.25, $p < 0.001$; interaction: $\beta$ = -0.03, $p = 0.66$). A comparison of generalized linear mixed-effects models predicting choice from the heuristic rules and general self-esteem (which fit the model to all participants simultaneously - an approach necessary for comparisons including general self-esteem, due to only one value per participant), again confirms that the reference-point heuristic based on average self-evaluation fits better to the data than other heuristics (BIC values for heuristics based on:

self-rating, 3725.2; self-rating – average other rating, 3721.6; self-rating – average self-rating, 3716.7).

## Individual differences in responses to surprising information are related to confirmatory information seeking

People who react more negatively to surprises should be more motivated to minimize them. To test if this is indeed the case, we estimated the strength of cognitive (confidence updates) and hedonic (momentary feeling) responses to surprises separately for each participant. To orthogonalize these responses, each estimate controlled for the other type of response. We find that the stronger the cognitive response to surprise (decrease in confidence following surprise), the more often a person was making confirmatory choices on average, as defined in the reference-point heuristic (partial correlation: R = -0.32, p = 0.01, controlling for the hedonic response to surprises, **Figure 5A**). Hedonic response to surprises was not significantly related to confirmatory choices (partial correlation: R = 0.04, p = 0.77, controlling for the cognitive response to surprises **Figure 5B**).

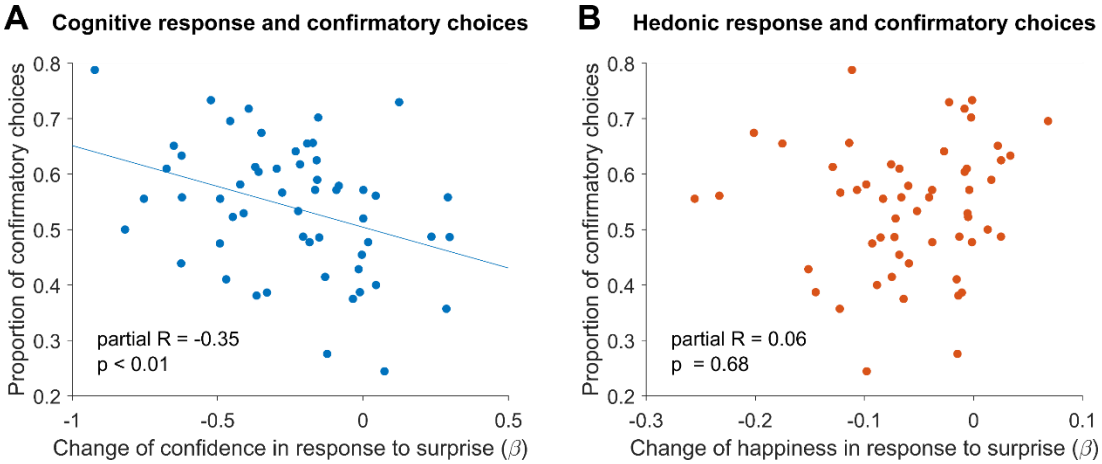


**Figure 5. Correlations between proportion of confirmatory choices and strength of responses to surprises.** For each participant, we computed the proportion of confirmatory choices, defined as confirmatory whenever choosing positive information for traits with self-ratings higher than average or choosing negative information for traits with self-rating lower than average (based on the reference-point heuristic). We also estimated the strength of

cognitive and hedonic response to surprises separately for each participant, while controlling for the other type of response. Plot (A) shows a significant partial correlation between cognitive response to surprises (confidence updated) and confirmatory choices, suggesting that the more negative the response, the more frequent confirmatory choices. Plot (B) shows that we did not find such relation for hedonic response (feelings ratings).

Another possible interpretation of the relation between the magnitude of cognitive response to surprises and the tendency to make confirmatory information choices is that making a confirmatory choice could enhance cognitive response to surprise. To test for this possibility, we created a GLME model predicting confidence update from experienced surprise, and an interaction of surprise with the type of choice: confirmatory or disconfirmatory (defined based on the average self-reference heuristic). We find that the cognitive response to surprise was the same, irrespective if it followed confirmatory or disconfirmatory choice (interaction: $\beta = 0.02$, $p = 0.83$). Lack of difference suggests that it is more likely that the relation between the magnitude of cognitive response to surprises and the tendency to make confirmatory choices is due to participants with stronger cognitive responses to surprises making more frequent confirmatory choices to avoid such surprises, rather than confirmatory choices causing a stronger response to surprises.

**Confidence updates show positivity bias**

We replicate previous findings (Korn et al., 2012; Koban et al., 2017), showing that people update their self-evaluations to a greater extent after receiving better than expected information than worse than expected information (positive prediction error: $\beta = 0.41$, 95% CI [0.33, 0.49] $p < 0.001$; negative prediction error: $\beta = 0.23$, 95% CI [0.20, 0.27] $p < 0.001$; difference between coefficients: $F(1) = 14.78$, $p < 0.001$). We extend the above findings by showing that people not only learn more from positive information than

negative information but also are more confident when they update their beliefs in more positive than negative direction (signed belief update: β = 0.21, p <0.001). There are two possible explanations of this phenomenon. As suggested by previous research, people with low self-esteem tend to have a less clear view of themselves (Baumgardner, 1990; Campbell, 1990). Therefore, any negative belief about oneself might be inherently uncertain. An alternative explanation would be that the learning process itself enhances confidence if it moves the belief in a positive direction, irrespective if of how positive or negative is the updated belief. To disentangle these two explanations, we created a GLME model predicting the confidence update from the valence of the final updated belief, the signed magnitude of the belief update, and the surprise about the received information. We find that all of these factors significantly predicted the updates of confidence (valence of final belief: β = 0.04, p = 0.04, belief update: β = 0.17, p < 0.001, surprise: β = -0.12, p < 0.001). The signed magnitude of the belief update was significantly more important in predicting the confidence update than how positive/negative was the final belief (belief update: β 95% CI [0.11, 0.24]),], final belief: β 95% CI [0.00, 0.08]; F(1) = 9.81, p < 0.01). These results suggest that updating one's beliefs in a positive direction, and to a lower extent holding a positive self-belief are related to increased confidence in these beliefs.

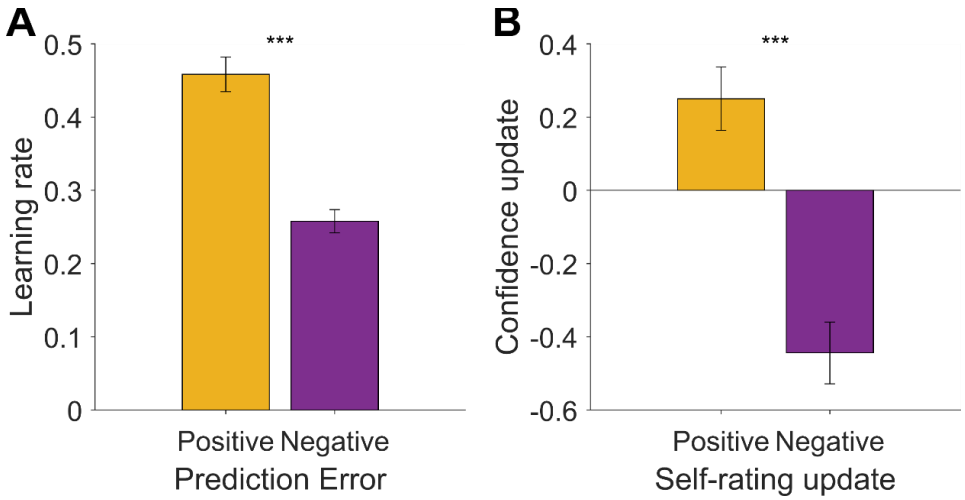


**Figure 6. Self-beliefs and confidence updates show positivity bias.** (A) To illustrate the positivity in updating beliefs about oneself we divided trials into trials where participants

experienced positive and negative prediction errors. We next calculated the average learning rate for each type of these trials for each participant, that describes how much a discrepancy between initial belief and revealed rating was incorporated in the new belief about oneself (see methods for details), and averaged them over participants. We find that participants updated their beliefs about oneself significantly more after experiencing positive prediction errors than negative prediction errors ($t(63) = 5.14$, $p < 0.001$). (B) To illustrate the positivity bias in updating confidence about specific self-evaluation, we divided trials into trials where participants updated their beliefs in the positive or negative direction. We next calculated the average confidence update for each type of these trials for each participant and averaged them over participants. We find that participants increased their confidence after updating their beliefs in the positive direction, but decreased their confidence after updating their beliefs in the negative direction ($t(63) = 5.50$, $p < 0.001$). These analyses are equivalent to the regression models described in the main manuscript. ***, $p < 0.001$.

## DISCUSSION

People often prefer to receive information that confirms their self-relevant beliefs from information that is flattering – a motive known as self-verification (Pettit and Joiner, 2001; Kwang and Swann, 2010; Kappes et al., 2020). Despite such preference being well-documented, there are many unanswered questions regarding the processes involved in such confirmatory information-seeking. How do we decide what information to seek when we are uncertain which source will confirm our beliefs, and what are the consequences of receiving (un)expected information about ourselves? What are the computations underlying these effects? To answer these questions, we designed an experiment in which people repeatedly faced an opportunity to reveal uncertain positive or negative ratings about their personality traits coming from other people. After revealing each rating, we prompted our participants to report how they felt at the current moment. Prior and after this part of the experiment, participants were asked to rate their personality traits and provide a confidence level in their rating. We find that receiving surprising information about oneself (controlling for other characteristics of feedback),

induces negative feelings and decreases one's confidence in the specific self-rating, providing a reason to avoid such surprises.

We tested several different choice algorithms that could serve such a goal, which can be grouped into general categories of heuristics and learning processes. Our results show that a heuristic that relies on a reference-point based on average self-rating outperformed all other alternatives. According to this choice strategy, participants will be more likely to choose positive sources of information the higher they rate their specific personality trait in comparison to how their rate their personality traits on average, and will more likely to choose negative source of information the lower they rate this trait in comparison to how they rate themselves on average. We find that this choice strategy is independent of the effect of general self-esteem on information preferences, suggesting the existence of at least two separate routes through which self-verification motive exhibits its influence on information choices.

The finding that surprising information about oneself induces negative hedonic response is an important one, as it challenges few commonly held assumptions. First assumption is related to the fact that many previous studies failed to find the effect of self-verification on affective measures, leading to a conclusion that self-verification primarily influences choice preferences rather than emotional reactions to outcomes (Kwang and Swann, 2010). A null finding made some researchers contest the idea that people care about consistency between information and beliefs about oneself and suggest other explanations for self-verifying choice patterns (Kruglanski et al., 2018). The reason why the current study was able to identify such an affective response is a methodological one. To the best of our knowledge, all previous experimental studies that investigated affective responses to self-verifying information focused on inter-individual differences between people with low and high general self-esteem, or a single self-evaluation related to one domain, for example appearance or performance beliefs (Swann et al., 1987; Ayduk et al., 2013; Jussim, Yen and Aiello, 1995; Moreland and Sweeney, 1984; Quinlivan

and Leary, 2005; Robinson and Smith-Loving, 1992; Ralph and Mineka, 1998; Wood et al., 2005; Stake, 1982), as opposed to within-individual discrepancies between a variety of specific self-views and received information as in the current study. These studies generally find that people have a negative emotional response to negative feedback, irrespective of their general self-esteem – a pattern more consistent with self-enhancement than self-verification motive. Our results replicate these findings, as we also did not find a significant influence of a discrepancy between received rating and general self-esteem on feelings. However, we do find a significant influence of a specific surprise, defined as an unsigned difference between received rating and self-rating on a specific trait: the more different the received rating from how the person rated themselves, the more unhappy the person was, irrespective of how good or bad was the rating itself. Furthermore, we find that the rating received and the signed prediction error (that is the difference between self-rating and received rating) had additional and independent effects on experienced feelings. The simultaneous effect of these three factors could be an additional reason why many studies failed to find an effect of self-verification on feelings, as disentangling these factors is only possible when the specific surprise about the outcome is precisely quantified.

A significant negative effect of specific surprise on emotional response also suggests that the current computational models of changes in momentary subjective well-being need to be augmented with additional factors (Will et al., 2017; Rutledege et al., 2014). In particular, our results imply that, all other things considered, subjects are happier if they receive information that is unsurprising rather than exceeding their expectations. The opposite prediction would be made if positive prediction errors were the sole factor driving momentary happiness. Future work will need to address in which contexts surprises have a negative hedonic value, and in which situations predicting happiness just based on signed prediction error is sufficient.

We also found that receiving surprising information about a trait resulted in a negative cognitive response, measured by a decrease of participants' confidence in their self-ratings about a trait. Despite an explicit assumption in self-verification theory that receiving inconsistent information with one's self-evaluations induces a state of uncertainty (Swann, 1983), there were almost no attempts to demonstrate this directly. One study that aimed to address this question investigated how receiving feedback consistent or inconsistent with one's general self-esteem affects the concept of self-clarity, defined as stability, internal consistency and confidence in one's general self-concept, and measured by a questionnaire (Stinson, Holmes, and Forest, 2010). Consistent with the self-verification theory, the study found that receiving inconsistent feedback with one's general self-esteem decreased self-clarity. In our study, a discrepancy between general self-esteem and ratings received from others did not significantly affect confidence updates. However, we did find a significant influence of a specific surprise on confidence updates: the bigger the discrepancy between specific self-rating and rating received from others in absolute terms, the more participants decreased their confidence in their self-rating on that specific trait. Future studies will need to address a question of how a concept of self-clarity is linked to confidence in specific self-evaluations. Differential result between distance from general self-esteem and distance from specific self-rating in our study suggests that the confidence question might measure a distinct phenomenon from the concept of general self-clarity.

Apart from the effect of surprises, we also find that confidence updates depend on updating one's self-ratings up or down: the more positive the self-rating update, the more people increase their confidence in their self-rating, and the more negative the rating update, the more people decrease their confidence. This finding suggests an existence of a positivity bias in confidence updates, complementing the positivity bias in belief updates about oneself (Korn et al., 2012; Koban et al., 2017). We disentangle two possible mechanisms

underlying this effect. According to one possibility, holding a positive belief in itself might be related to high confidence, as suggested by studies showing that people with low self-esteem tend to have a lower concept of self-clarity (Baumgardner, 1990; Campbell, 1990). Such relation might be adaptive due to the ease of modifying an uncertain negative belief or a lower hedonic burden of it. Another possibility could be that updating a belief in positive direction enhances confidence, irrespective of how positive or negative is the final belief. Our analysis supports the existence of both of these mechanisms.

The negative hedonic and cognitive utility of surprising information about oneself provides a reason for people to avoid such surprises. However, do people take into account these factors when seeking information? Although our study does not provide a direct test of this hypothesis, we observed individual differences between participants that do provide some clues about the importance of these two responses in self-verification. In particular, we see that participants who decreased their confidence in their ratings more strongly after experiencing surprises were also more frequently making confirmatory choices aligned with the reference-point heuristic outlined above. We did not find a similar relation for a hedonic response to surprises, supporting the idea that cognitive motives might primarily drive self-verification. As this is a correlational result, the directionality of influence is uncertain. However, we did not find a significant difference between the cognitive reaction to surprises after making a confirmatory and disconfirmatory choice, suggesting that it is more likely that participants who experience a more negative cognitive reaction to surprises might try to avoid such surprises by making confirmatory choices, rather than confirmatory choice enhancing a reaction to surprises.

If people are motivated to avoid surprises, how do they decide what information source to choose to achieve this goal? Information by definition has uncertain content, as opposed to feedback, that is orthogonal to uncertainty. Therefore, agents face a problem of what decision rule to apply to

minimize possible surprises when consequences of actions are uncertain and different strategies yield a different probability of success. We tested the optimality of several different heuristics and learning mechanisms that aim to minimize surprise in a simulated environment similar to the one used in our experiment. In this environment, artificial agents had to repeatedly sample between two sources of information, knowing only that one is on average more positive than the other, but not knowing anything else about the distribution of information in each source. We show that a heuristic that always chooses positive information source whenever a specific belief is higher than the average belief about oneself, and negative information whenever a specific belief is lower than average belief about oneself, is the most optimal strategy for minimizing surprises if the average belief about oneself is close to an average received information from the two sources. Such an assumption is reasonable in the real world, as people might expect that the received information about oneself should be on average close to what a person believes to be true about oneself and consistent with the notion that heuristics often represent optimal solutions to everyday problems. The more average belief about oneself deviates from the average from information sources, the bigger the advantage of learning processes, especially model-based algorithms that learn the structure of the environment and simulate possible outcomes of actions before making a choice. The competitive advantage of learning processes also increases the more apart are the two information sources due to the increased cost of making an 'erroneous' response in terms of surprise.

Having identified which strategy is the most efficient in minimizing surprises, next we evaluated which strategy fits best to participants' behavior. Overall, a reference-point heuristic based on average self-rating outperformed all other models. However, we observed substantial heterogeneity between individuals. In 44.6% of cases participants' behavior was best characterized by a learning mechanism. Based on our simulation, we know that participants

would benefit the most from using a learning strategy if their general self-esteem deviated from the average rating that they could expect from the information sources, and should use a heuristic rule otherwise. Indeed, we observe that the general self-esteem of participants who use learning mechanisms deviated from the average rating in the task significantly more than the general self-esteem of participants using heuristics. At the same time, we were not able to identify any other significant difference between these two groups. This suggests that participants are able to choose a strategy that is the most optimal in minimizing surprises given their specific circumstances. Future work will need to identify if all participants start with a heuristic and move to a learning strategy once it proves insufficient in minimizing surprises, or perhaps use other methods of identifying the best course of action.

Variability in used strategies in this task might be stemming from ambiguity about the extent to which received feedback was predictive of future feedback. If participants assumed that each personality rating was done independently, without any correlation between separate ratings, then there would be no point in learning anything. On the other hand, if participants assumed that ratings are correlated with each other, then learning about the average rating would allow them to develop more sophisticated strategies of minimizing surprises.

We find that the usage of reference-point heuristic affects the logarithm of the reaction time of choices. In particular, we find that reaction time is inversely related to how strongly a participant is influenced by the reference-point heuristic in their choices, consistent with the suggestion that usage of heuristics allows fast decisions. Additionally, we find that the probability of choosing positive information, estimated based on reference-point heuristic, is inversely related to the logarithm of reaction time. That is, the higher the person's self-rating in comparison to their average self-rating, the faster the decision, suggesting that despite lack of clear preferences for positive over negative information in our experiment, participants still exhibit positivity bias

in their reaction time – similar to other studies (Leppänen, Tenhunen and Hietanen, 2003; Guitart-Masip, et al., 2011; Gesiarz, Cahill, and Sharot, 2019). At the same time, we did not find a significant difference in reaction times between positive and negative choices of information, suggesting that taking into account the reference-point heuristic might be necessary to uncover such bias.

Our findings create a link between so far separate lines of research and provide a framework for future studies that aim to look at self-verification behavior from a computational perspective. By identifying hedonic and cognitive responses to surprise signals accompanying self-confirmatory choices, and strategies that people use to minimize these surprises, we validate and extend the self-verification theory. There are many outstanding topics that this framework might contribute to, including our understanding of processes involved in confirmation biases in general, and the role of self-verification in major depression disorder and social anxiety, in which negative self-image is reinforced by seeking negative information about oneself (Valentiner et al., 2011; Giesler, Josephs, and Swann, 1996; Joiner, Katz and Lew, 1997). In a broader context, the current study helps in advancing our knowledge about complex motives driving information-seeking, that constitute an increasingly important part of personal and professional life.

# METHODS

## Participants

We recruited 94 participants from University College London subject to take part in our experiment. We excluded from the analysis 29 participants who did not believe that the experimental manipulation was true in a debriefing questionnaire, resulting in a sample of 65 people (of which were 48 female, mean age 23.5). Participants originated from 22 different countries, and 48% of them originated from countries classified as western. All participants

provided written informed consent. The experiment was approved by the UCL ethics committee.

**Procedure**

***Overview.*** We invited participants to the lab in groups of four/three. Volunteers were remunerated for their time at the rate of £7.50 per hour. The experiment had four parts: (1) getting to know the other participants, (2) evaluating oneself and others, (3) choosing information about their ratings, (4) re-evaluation of oneself. After the experiment, participants filled in a demographics questionnaire and were provided with debriefing about the study.

***Getting to know others.*** Participants were paired with one other person, asked to introduce himself or herself and chat for 10 minutes about three provided topics with each other. The topics included: (a) "what would constitute a perfect day for you?", (b) "what would you like to be famous for?", (c) "if you could wake up tomorrow having gained any one quality or ability, what would it be?", and were taken from a list of topics used for induction of interpersonal closeness (Aron et al., 1997). After 10 minutes, participants were paired with another person and asked again to chat about the same topics. This part of the experiment aimed to provide broader familiarity with two newly met people that would allow participants to subsequently evaluate their impression of others.

***Self and other evaluations.*** After getting to know other each other, participants were directed to separate cubicles and asked to rate themselves and others on a list of 50 positively valenced adjectives. The adjectives were taken from a questionnaire that mapped the Big-5 traits to English adjectives, and were displayed at the top of the screen with a short definition taken from a dictionary (Goldberg, 1992). The provided scale ranged from 0 to 10, with 0

corresponding to a trait being not at all characteristic for a person, and 10 being very much characteristic for a person. To avoid anchoring effects of a starting position of the marker, participants had to click on the unmarked scale to provide their rating. First, participants had to rate themselves, and also provide a confidence level in their self-evaluations, on a scale ranging from 0 to 10, with 0 corresponding to being very uncertain about the provided rating, and 10 to being very certain about the provided rating. Subsequently, they were asked to rate the first person that they talked to on the list of the same adjectives, as well as the second person that they talked. After this part, participants had to wait for all participants to finish their ratings.

***Information choice.*** Participants were informed that they will have now an opportunity to find out how they were rated by others. They were provided with a choice: either they could reveal a higher rating that they received or the lower rating that they received for an adjective displayed at the top of the screen. The adjectives were presented in a random order. The colours marking higher and lower option were counterbalanced between participants. There were 50 choice trials in total. After the choice, the rating was displayed on the screen for 2 seconds. To ensure that participants processed the information, they had to wait 5 seconds for the next trial. After 5 seconds, they were asked about how they are currently feeling. Participants provided an answer by clicking on a continuous scale ranging very happy to very unhappy.

Unbeknownst to participants, all revealed ratings were generated by a computer, which in each trial drew two numbers from a normal distribution, with a mean of 6 and standard deviation of 2, rounded to the nearest integer, and sorted them so the lower number was assigned to the negative rating and higher number to the positive rating (or randomly assigned in the case when the numbers were equal). If the drawn number was greater than 10 or lower than 3, it was equalized to the bound that it exceeded. On average, people

experienced signed prediction error equal to -1.14(1.19), and surprise equal to 2.26(0.56).

***Self re-evaluations.*** In the last part, participants were asked to again rate themselves on a list of provided adjectives, that were presented in the same form and order as during the first time participants were evaluating themselves. They were also asked about their confidence in the rating.

**Data analysis**

All analysis was performed using MATLAB 2019a software.

***Dependent variables.*** Choices were coded as 0 and 1, with 0 corresponding to a decision to reveal lower rating (negative information) and 1 as decision to reveal higher rating (positive information). Feelings ratings were recoded to range from 0 to 1, with 0 corresponding to very unhappy and 1 to very happy. Rating update was defined as a difference between second self-evaluation and first self-evaluation, separately for each adjective. Confidence update was defined as a difference between confidence rating during second self-evaluation and first self-evaluation, separately for each adjective.

***Independent variables.***

Surprise was defined as absolute (unsigned) difference between revealed and first self-rating for each trial:

$$\text{surprise} = |\text{revealed rating}_t - \text{self rating}_t|$$

Prediction error was defined as a signed difference between revealed and first self-rating for each trial:

$$\text{prediction error} = \text{revealed rating}_t - \text{self rating}_t$$

Positive prediction errors were defined as:

$$\text{positive prediction error} =$$

$$\begin{cases} \text{revealed rating}_t - \text{self rating}_t, \text{if revealed rating}_t - \text{self rating}_t > 0 \\ 0, \text{ if revealed rating}_t - \text{self rating}_t \leq 0 \end{cases}$$

Negative prediction errors were defined as:

$$\text{negative prediction error} =$$

$$\begin{cases} \text{revealed rating}_t - \text{self rating}_t, \text{if revealed rating}_t - \text{self rating}_t < 0 \\ 0, \text{ if revealed rating}_t - \text{self rating}_t \geq 0 \end{cases}$$

Value of the received information was assumed to be proportional to the revealed rating.

General self-esteem was calculated as an average from all self-ratings:

$$\text{general self esteem} = \sum_{t=1}^{n} \text{self rating}_t$$

General surprise was calculated as follows:

$$\text{general surprise} = |\text{revealed rating}_t - \text{general self} - \text{esteem}_i|$$

Where *i* is participant's index.

General prediction error was calculated as follows:

$$general\ prediction\ error = revealed\ rating_t - general\ self - esteem_i$$

***Generalized linear mixed-effects model.*** To account for within-subject correlations of responses related to repeated measures in our design, we used Generalized Linear Mixed Effects (GLME) model approach, in which fixed effects describe effects common for all participants and random effects describe idiosyncrasies specific for an individual (Bar et al., 2013). All models included intercept and a random effect for each fixed effect, as recommended by Bar and colleagues (2013). To obtain standardized beta coefficients, all independent variables were z-scored prior to the analysis. To compare magnitude of coefficients, we used 95% confidence intervals and to obtain a p-value for this comparison we used an F-test.

To analyse the impact of received information on cognitive and hedonic responses we used the following GLME models:

$$\text{Confidence update}_t = \beta_0 + \beta_1 \text{prediction error}_t + \beta_2 \text{surprise}_t + \beta_3 \text{received rating}_t$$

$$\text{Feelings ratings}_t = \beta_0 + \beta_1 \text{prediction error}_t + \beta_2 \text{surprise}_t + \beta_3 \text{information valence}_t$$

To analyse the potential impact of received information on self-clarity and accompanying it cognitive and hedonic responses we used the following GLME models:

$$\text{Confidence update}_t = \beta_0 + \beta_1 \text{general prediction error}_t + \beta_2 \text{general surprise}_t + \beta_3 \text{received rating}_t$$

$$\text{Feelings ratings}_t = \beta_0 + \beta_1 \textit{general} \text{ prediction error}_t + \beta_2 \text{general surprise}_t + \beta_3 \text{information valence}_t$$

To analyse positivity bias, we used the following GLME models:

$$\text{Ratings update}_t = \beta_0 + \beta_1 \text{positive prediction error}_t + \beta_2 \text{negative prediction error}_t$$

$$\text{Confidence update}_t = \beta_0 + \beta_1 \text{surprise}_t + \beta_2 \text{ratings update}_t + \beta_3 \text{second self evaluation}_t$$

***Between group and within-participants categorical comparisons.*** All between group and within-participants comparisons based on categorical classification of trials were performed using a two-tailed independent and dependent samples t-test respectively. In many cases they duplicate the above GLME model analysis performed on continuous variables, and serve illustratory purposes (**Figure 2**, **Figure 3F**, and **Figure 5**).

***Choice mechanisms analysis.*** We compared several possible mechanisms driving information choice. Each model was fit separately to individual's behaviour, using fmnincon function in MATLAB. To minimize the chance of finding a local rather than global minimum, the fitting procedure was repeated

100 times, each time with different random starting values for parameters. To compare overall model fit and parsimony we used Bayesian Information Criterion, summed over all participants (Smith and Spiegelhalter, 1980). Individual fits computed for all models and all participants allowed us to additionally calculate percentage of participants for which each model was the best model out of all models minimizing surprises.

Absolute scale heuristic assumed that people have a tendency to reveal higher option whenever their self-rating is higher than the middle of the scale:

$$\text{choice value}_t = \text{self rating}_t$$

Average self-evaluation as a reference point heuristic assumes that people have a tendency to reveal higher option whenever their self-rating is higher than their average self-evaluation:

$$\text{choice value}_t = (\text{self rating}_t - \text{general self esteem})$$

Average evaluation of others as a reference point heuristic assumes that people have a tendency to reveal higher option whenever their self-rating is higher than their average evaluation of others on that specific trait:

$$\text{choice value}_t = \text{self rating}_t - \left(\frac{\text{other 1 rating}_t + \text{other 2 rating}_t}{2}\right) + \text{intercept}$$

The choice was modelled as a softmax function:

$$\text{probability of revealing positive information} = \frac{1}{1 + e^{-(\beta_1 \text{choice value}_t + \beta_0)}}$$

With $\beta_1$ being an inverse temperature parameter, modifying the sensitivity to change in choice value, and $\beta_0$ being an intercept.

Learning processes were modelled as follows.

*Model-free learning* assumes that people learn to avoid choices that lead to a surprise in the past.

$$\text{higher value}_{t+1} = \begin{cases} \text{higher value}_t + \alpha(\text{surprise}_t - \text{higher value}_t), \text{if higher option was chosen} \\ \text{higher value}_t, \text{if lower option was chosen} \end{cases}$$

$$\text{lower value}_{t+1} = \begin{cases} \text{higher value}_t + \alpha(\text{surprise}_t - \text{lower value}_t), \text{if lower option was chosen} \\ \text{lower value}_t, \text{if higher option was chosen} \end{cases}$$

$$\text{probability of revealing positive information} = \frac{1}{1 + e^{-(\beta_1(\text{lower value}_t - \text{higher value}_t) + \beta_0)}}$$

With higher and lower values of choices initialized to 0.

*Model-based learning* assumes that people learn the average rating that they can expect after revealing higher or lower option, and choose the option that they expect to minimize surprise.

$$\text{higher average}_{t+1} = \begin{cases} \text{higher average}_t + \alpha(\text{revealed rating}_t - \text{higher average}_t), \text{if higher option was chosen} \\ \text{higher average}_t, \text{if lower option was chosen} \end{cases}$$

$$\text{lower average}_{t+1} = \begin{cases} \text{lower average}_t + \alpha(\text{revealed rating}_t - \text{lower average}_t), \text{if lower option was chosen} \\ \text{lower average}_t, \text{if higher option was chosen} \end{cases}$$

$$\text{model based higher value}_t = |\text{self rating}_t - \text{higher average}_t|$$

$$\text{model based lower value}_t = |\text{self rating}_t - \text{lower average}_t|$$

$$\text{probability of revealing positive information} = \frac{1}{1 + e^{-(\beta_1(\text{model based lower value}_t - \text{model based higher value}_t) + \beta_0)}}$$

With higher and lower average expected ratings initialized to general self-esteem of the participant.

*Hybrid model-based and model-free learning model* assumes that people use both strategies to some extent.

$$\text{probability of revealing positive information} = \frac{1}{1 + e^{-(\beta_1(\text{model based lower value}_t - \text{model based higher value}_t) + \beta_2(\text{lower value}_t - \text{higher value}_t) + \beta_0)}}$$

With $\beta_1$ and $\beta_2$ controlling the balance between model-based and model-free strategies respectively.

We additionally tested learning processes that aimed to maximize positive prediction errors, rather than minimize surprises, specified as follows.

*Model-free learning:*

$$\text{higher value}_{t+1} = \begin{cases} \text{higher value}_t + \alpha(\text{prediction error}_t - \text{higher value}_t), \text{if higher option was chosen} \\ \text{higher value}_t, \text{if lower option was chosen} \end{cases}$$

$$\text{lower value}_{t+1} = \begin{cases} \text{higher value}_t + \alpha(\text{prediction error}_t - \text{lower value}_t), \text{if lower option was chosen} \\ \text{lower value}_t, \text{if higher option was chosen} \end{cases}$$

$$\text{probability of revealing positive information} = \frac{1}{1 + e^{-(\beta_1(\text{higher value}_t - \text{lower value}_t) + \beta_0)}}$$

With higher and lower values of choices initialized to 0.

Model-based learning:

$$\text{higher average}_{t+1} = \begin{cases} \text{higher average}_t + \alpha(\text{revealed rating}_t - \text{higher average}_t), \text{if higher option was chosen} \\ \text{higher average}_t, \text{if lower option was chosen} \end{cases}$$

$$\text{lower average}_{t+1} = \begin{cases} \text{lower average}_t + \alpha(\text{revealed rating}_t - \text{lower average}_t), \text{if lower option was chosen} \\ \text{lower average}_t, \text{if higher option was chosen} \end{cases}$$

$$\text{model based higher value}_t = \text{self rating}_t - \text{higher average}_t$$

$$\text{model based lower value}_t = \text{self rating}_t - \text{lower average}_t$$

$$\text{probability of revealing positive information} = \frac{1}{1 + e^{-(\beta_1(\text{model based higher value}_t - \text{model based lower value}_t) + \beta_0)}}$$

With higher and lower average expected ratings initialized to general self-esteem of the participant.

*Hybrid model-based and model-free learning model:*

$$= \frac{1}{1 + e^{-(\beta_1(\text{model based higher value}_t - \text{model based lower value}_t) + \beta_2(\text{lower higher}_t - \text{higher lower}_t) + \beta_0)}}$$

***Simulation.*** To investigate the optimality of different choice strategies that aim to minimize surprises, we created simulated agents that used either a reference-point heuristic based on average self-esteem, model-based learning or model free-learning in their choices, in an environment similar to the one used in our task. The reference-point heuristic based on average evaluation of others was not modelled due to additional complexity of modelling the relation between self-evaluation and evaluation of others. The agents could be characterized by different level of general-self esteem, ranging from deviating -5 from the average expected rating from the two sources of information to +5. Their specific self-evaluations on 50 traits were normally distributed around this general self-esteem, with a standard deviation equal to 2. Each agent started with no knowledge of the environment, except a knowledge that one source of information is more positive than the other.

Each source of information provided a rating drawn from a normal distribution, with a mean $u$ and standard deviation of 2. The mean rating of each source varied between simulations. In the example in the main manuscript it was equal to either 4, 3, or 2 for the negative information source and either 6, 7, or 8 for the positive information source (for **Figure 3A**, **3B**, and **3C** respectively). The example in the manuscript presents a situation where information sources are unreliably positive and negative, i.e. on average positive source of information presents higher ratings than negative information sources, but is not guaranteed to always present a better rating in every instance. Supplementary material presents additional simulation where the positive information source is reliably positive.

We created 2000 agents per each algorithm and general self-esteem level. Each agent's choice was fully deterministic, that is whenever the probability of positive choice exceeded 50% they always chose positive information source. The learning rate parameter was set to 0.3 in the example in the main

manuscript. Supplementary Material additionally presents simulations with learning rates set to 0.1 and 0.5. All other specifications of the models were applied as described above.

Each agent made 50 decisions. We then averaged the experienced surprise during the task over 50 trials, and averaged over agents separately for each algorithm and general self-esteem level.

***Comparisons of GLME heuristics choice models including general self-esteem.*** Incorporating general self-esteem into individual fits is not possible, due to only a single value for each participant. Therefore, too compare different heuristics, while including general self-esteem, we used the following GLME models:

$$\text{Choice}_t = \beta_0 + \beta_1 self\ rating + \beta_2 \text{general self esteem}_t$$

$$\text{Choice}_t = \beta_0 + \beta_1 self\ reference - point\ heuristic + \beta_2 \text{general self esteem}_t$$

$$\text{Choice}_t = \beta_0 + \beta_1 other\ reference - point\ heuristic + \beta_2 \text{general self esteem}_t$$

Where $\beta_0$ is the intercept, and *self reference-point heuristic* is the difference between participant's self-rating and average self-evaluation, and *other reference-point heuristic*, is the difference between participant's self-rating and average evaluation of others. We used Bayesian Information Criterion to compare the fit of these models.

Additionally, we run the following GLME model to test the independence reference-point heuristic from the effect of general self-esteem

$$\text{Choice}_t = \beta_0 + \beta_1 self\ reference - point\ heuristic_t + \beta_2 \text{general self esteem}_t + \beta_3 self\ reference - point\ heuristic_t * \text{general self esteem}_t$$

***Confirmatory choices.*** Confirmatory choices were defined based on the winning choice heuristic: the choice was confirmatory if a person chose to reveal higher rating, if their self-rating was higher than general self-esteem for that specific trait, or chose to reveal lower rating if their self-rating was lower

than general self-esteem for that specific trait. The choice was classified as disconfirmatory otherwise.

***Analysis of reaction times.*** The decision reaction time was log-transformed due to its' highly skewed distribution – a procedure common for analysis of reaction times (Lo and Andrews, 2015). We used to following GLME model to estimate the effect of positive/negative and confirmatory information choices on rection time (RT):

$$\text{Log RT}_t = \beta_0 + \beta_1 confirmatory\ choice_t + \beta_2 \text{positive/negative choice}_t$$

Next we investigated if reaction times can be predicted based on the reference-point model. For that purpose, we extracted the estimated probability of revealing positive information for each trial and participant based on the reference-point heuristic. We tested two possibilities. One being that reaction time is linearly related to choice probability, and second that reaction time is related to choice uncertainty, following an inverted U-shape peaking at 50%, and quantified as follows:

$$choice\ uncertainty_{\ t} = P(revealing\ positive\ information_t) * (1 - P(revealing\ positive\ information_t))$$

Where $P(revealing\ positive\ information_t)$ is the probability of choosing positive information choice, and $t$ is the trial number.

We used to following GLME model to estimate the effect of positive information choice and choice uncertainty on reaction times:

$$\text{Log RT}_t = \beta_0 + \beta_1 choice\ uncertainty_{\ t} + \beta_2 P(revealing\ positive\ information_t)$$

We also tested to if reliance on the reference-point heuristic influenced reaction times in general. To test that, we extracted from fitted reference-point heuristic model the $\beta_1$ parameter, which describes how much person's choices are influenced by the difference between specific self-rating and general self-

esteem. We next correlated these $\beta_1$ estimates with mean reaction time of each participant.

***Analysis of individual differences in responses to surprise.*** To analyse the individual differences in a relation between the strength of responses to surprises and individual's behaviour, we estimated the strength of hedonic and cognitive responses for each individual separately, while controlling for the other type of response.

$$\text{Cognitive response}_t = \beta_0 + \beta_1 \text{surprise}_t + \beta_2 \text{Feelings ratings}_t$$
$$\text{Hedonic response}_t = \beta_0 + \beta_1 \text{surprise}_t + \beta_2 \text{Cogntive response}_t$$

We next performed a partial correlation of $\beta_1$ for hedonic and cognitive responses with an average proportion of confirmatory choices of each participant, controlling for the other type of response.

## REFERENCES


Aron, A., Melinat, E., Aron, E. N., Vallone, R. D., and Bator, R. J. (1997). The Experimental Generation of Interpersonal Closeness: A Procedure and Some Preliminary Findings. *Personality and Social Psychology Bulletin*, *23*(4), 363–377. https://doi.org/10.1177/0146167297234003

Ayduk, Ö., Gyurak, A., Akinola, M., and Mendes, W. B. (2013). Consistency Over Flattery: Self-Verification Processes Revealed in Implicit and Behavioral Responses to Feedback. *Social Psychological and Personality Science*, *4*(5), 538–545. https://doi.org/10.1177/1948550612471827

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Baumeister, R. F., and Hutton, D. G. (1987). Self-Presentation Theory: Self-Construction and Audience Pleasing. In B. Mullen and G. R. Goethals (Eds.), *Theories of Group Behavior* (pp. 71–87). https://doi.org/10.1007/978-1-4612-4634-3_4

Baumgardner, A. H. (1990). To know oneself is to like oneself: self-certainty and self-affect. *Journal of Personality and Social Psychology*, *58*(6), 1062–1072. https://doi.org/10.1037//0022-3514.58.6.1062

Bernichon, T., Cook, K. E., and Brown, J. D. (2003). Seeking self-evaulative feedback: The interactive role of global self-esteem and specific self-views. *Journal of Personality and Social Psychology*, *84*(1), 194–204. https://doi.org/10.1037/0022-3514.84.1.194

Blaine, B., and Crocker, J. (1993). Self-Esteem and Self-Serving Biases in Reactions to Positive and Negative Events: An Integrative Review. In R. F. Baumeister (Ed.), *Self-Esteem: The Puzzle of Low Self-Regard* (pp. 55–85). https://doi.org/10.1007/978-1-4684-8956-9_4

Campbell, J. D. (1990). Self-esteem and clarity of the self-concept. *Journal of Personality and Social Psychology*, *59*(3), 538–549. https://doi.org/10.1037/0022-3514.59.3.538

Chen, S., English, T., and Peng, K. (2006). Self-Verification and Contextualized Self-Views. *Personality and Social Psychology Bulletin*, *32*(7), 930–942. https://doi.org/10.1177/0146167206287539

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., and Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*(6), 1204–1215. https://doi.org/10.1016/j.neuron.2011.02.027

Daw, N. D., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*(12), 1704–1711. https://doi.org/10.1038/nn1560

Dutton, K. A., and Brown, J. D. (1997). Global self-esteem and specific self-views as determinants of people's reactions to success and failure. *Journal of Personality and Social Psychology*, *73*(1), 139–148. https://doi.org/10.1037/0022-3514.73.1.139

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., O□Doherty, J., and Pezzulo, G. (2016). Active inference and learning. *Neuroscience and Biobehavioral Reviews*, *68*, 862–879. https://doi.org/10.1016/j.neubiorev.2016.06.022

Friston, K. J. (2018). Active Inference and Cognitive Consistency. *Psychological Inquiry*, *29*(2), 67–73. https://doi.org/10.1080/1047840X.2018.1480693

Friston, K., Schwartenbeck, P., Fitzgerald, T., Moutoussis, M., Behrens, T., and Dolan, R. J. (2013). The anatomy of choice: active inference and agency. *Frontiers in Human Neuroscience*, *7*. https://doi.org/10.3389/fnhum.2013.00598

Gesiarz, F., Cahill, D., and Sharot, T. (2019). Evidence accumulation is biased by motivation: A computational account. *PLOS Computational Biology*, *15*(6), e1007089. https://doi.org/10.1371/journal.pcbi.1007089

Gęsiarz, F., and Crockett, M. J. (2015). Goal-directed, habitual and Pavlovian prosocial behavior. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00135

Giesler, R. B., Josephs, R. A., and Swann Jr., W. B. (1996). Self-verification in clinical depression: The desire for negative evaluation. *Journal of Abnormal Psychology*, *105*(3), 358–368. https://doi.org/10.1037/0021-843X.105.3.358

Gigerenzer, G., Todd, P. M., and Group, A. R. (2000). *Simple Heuristics that Make Us Smart*. Oxford University Press.

Gläscher, J., Daw, N., Dayan, P., and O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595. https://doi.org/10.1016/j.neuron.2010.04.016

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Guitart-Masip, M., Fuentemilla, L., Bach, D. R., Huys, Q. J. M., Dayan, P., Dolan, R. J., and Duzel, E. (2011). Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain. *The Journal of Neuroscience*, *31*(21), 7867–7875. https://doi.org/10.1523/JNEUROSCI.6376-10.2011

Hong, S. L., and Beck, M. R. (2010). Uncertainty Compensation in Human Attention: Evidence from Response Times and Fixation Durations. *PLOS ONE*, *5*(7), e11461. https://doi.org/10.1371/journal.pone.0011461

Joiner Jr., T. E., Katz, J., and Lew, A. S. (1997). Self-verification and depression among youth psychiatric inpatients. *Journal of Abnormal Psychology*, *106*(4), 608–618. https://doi.org/10.1037/0021-843X.106.4.608

Jussim, L., Yen, H., and Aiello, J. R. (1995). Self-Consistency, Self-Enhancement, and Accuracy in Reactions to Feedback. *Journal of Experimental Social Psychology*, *31*(4), 322–356. https://doi.org/10.1006/jesp.1995.1015

Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., and Sharot, T. (2020). Confirmation bias in the utilization of others' opinion strength. *Nature Neuroscience*, *23*(1), 130–137. https://doi.org/10.1038/s41593-019-0549-2

Koban, L., Schneider, R., Ashar, Y. K., Andrews-Hanna, J. R., Landy, L., Moscovitch, D. A., … Arch, J. J. (2017). Social anxiety is characterized by biased learning about performance and the self. *Emotion*, *17*(8), 1144–1155. https://doi.org/10.1037/emo0000296

Koop, G. J., and Johnson, J. G. (2012). The use of multiple reference points in risky decision making. *Journal of Behavioral Decision Making*, *25*(1), 49–62. https://doi.org/10.1002/bdm.713

Korn, C. W., Prehn, K., Park, S. Q., Walter, H., and Heekeren, H. R. (2012). Positively Biased Processing of Self-Relevant Social Feedback. *Journal of Neuroscience*, *32*(47), 16832–16844. https://doi.org/10.1523/JNEUROSCI.3016-12.2012

Kruglanski, A. W., Jasko, K., Milyavsky, M., Chernikova, M., Webber, D., Pierro, A., and Santo, D. di. (2018). Cognitive Consistency Theory in Social Psychology: A Paradigm Reconsidered. *Psychological Inquiry*, *29*(2), 45–59. https://doi.org/10.1080/1047840X.2018.1480619

Kwang, T., and Swann, W. B. (2010). Do people embrace praise even when they feel unworthy? A review of critical tests of self-enhancement versus self-verification. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology, Inc*, *14*(3), 263–280. https://doi.org/10.1177/1088868310365876

Leary, M. R. (2007). Motivational and Emotional Aspects of the Self. *Annual Review of Psychology*, *58*(1), 317–344. https://doi.org/10.1146/annurev.psych.58.110405.085658

Leppänen, J. M., Tenhunen, M., and Hietanen, J. K. (2003). Faster choice-reaction times to positive than to negative facial expressions: The role of

cognitive and motor processes. *Journal of Psychophysiology*, *17*(3), 113–123. https://doi.org/10.1027//0269-8803.17.3.113

Lo, S., and Andrews, S. (2015). To transform or not to transform: using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*, 1171. https://doi.org/10.3389/fpsyg.2015.01171

Moreland, R. L., and Sweeney, P. D. (1984). Self-expectancies and reactions to evaluations of personal performance. *Journal of Personality*, *52*(2), 156–176. https://doi.org/10.1111/j.1467-6494.1984.tb00350.x

Pettit, J. W., and Joiner, T. E. (2001). Negative Life Events Predict Negative Feedback Seeking as a Function of Impact on Self-Esteem. *Cognitive Therapy and Research*, *25*(6), 733–741. https://doi.org/10.1023/A:1012919306708

Quinlivan, E., and Leary, M. R. (2005). Women's Perceptions of Their Bodies: Discrepancies Between Self-Appraisals and Reflected Appraisals. *Journal of Social and Clinical Psychology*, *24*(8), 1139–1163. https://doi.org/10.1521/jscp.2005.24.8.1139

Ralph, J. A., and Mineka, S. (1998). Attributional style and self-esteem: the prediction of emotional distress following a midterm exam. *Journal of Abnormal Psychology*, *107*(2), 203–215. https://doi.org/10.1037//0021-843x.107.2.203

Robinson, D. T., and Smith-Lovin, L. (1992). Selective Interaction as a Strategy for Identity Maintenance: An Affect Control Model. *Social Psychology Quarterly*, *55*(1), 12–28. https://doi.org/10.2307/2786683

Rosenberg, M., Schooler, C., Schoenbach, C., and Rosenberg, F. (1995). Global Self-Esteem and Specific Self-Esteem: Different Concepts, Different Outcomes. *American Sociological Review*, *60*(1), 141–156. https://doi.org/10.2307/2096350

Smith, A. F. M., and Spiegelhalter, D. J. (1980). Bayes Factors and Choice Criteria for Linear Models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*(2), 213–220. https://doi.org/10.1111/j.2517-6161.1980.tb01122.x

Stake, J. E. (1982). Reactions to positive and negative feedback: Enhancement and consistency effects. *Social Behavior and Personality*, *10*(2), 151-156–156. https://doi.org/10.2224/sbp.1982.10.2.151

Stinson, D. A., Logel, C., Holmes, J. G., Wood, J. V., Forest, A. L., Gaucher, D., … Kath, J. (2010). The regulatory function of self-esteem: testing the epistemic and acceptance signaling systems. *Journal of Personality and Social Psychology*, *99*(6), 993–1013. https://doi.org/10.1037/a0020310

Sutton, R. S., Barto, A. G., Barto, C.-D. A. L. L. A. G., and Bach, F. (1998). *Reinforcement Learning: An Introduction*. MIT Press.

Swann, W. (1983). Self-Verification: Bringing Social Reality into Harmony with the Self. In *Social psychological perspectives on the self* (Vol. 2, pp. 33–66). Retrieved from http://homepage.psy.utexas.edu/homepage/faculty/swann/docu/swBSRHS83.pdf

Swann, W. B., Griffin, J. J., Predmore, S. C., and Gaines, B. (1987). The cognitive-affective crossfire: when self-consistency confronts self-enhancement. *Journal of Personality and Social Psychology*, *52*(5), 881–889. https://doi.org/10.1037//0022-3514.52.5.881

Swann, W. B., Pelham, B. W., and Krull, D. S. (1989). Agreeable fancy or disagreeable truth? Reconciling self-enhancement and self-verification. *Journal of Personality and Social Psychology*, *57*(5), 782–791. https://doi.org/10.1037//0022-3514.57.5.782

Tesser, A., and Paulhus, D. (1983). The definition of self: Private and public self-evaluation management strategies. *Journal of Personality and Social Psychology*, *44*(4), 672–682. https://doi.org/10.1037/0022-3514.44.4.672

Tversky, A., and Kahneman, D. (1992). Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, *5*(4), 297–323. https://doi.org/10.1007/BF00122574

Valentiner, D. P., Skowronski, J. J., McGrath, P. B., Smith, S. A., and Renner, K. A. (2011). Self-Verification and Social Anxiety: Preference for Negative Social Feedback and Low Social Self-Esteem. *Behavioural and Cognitive Psychotherapy*, *39*(5), 601–617. https://doi.org/10.1017/S1352465811000300

Wang, X. T., and Johnson, J. G. (2012). A tri-reference point theory of decision making under risk. *Journal of Experimental Psychology: General*, *141*(4), 743–756. https://doi.org/10.1037/a0027415

Will, G.-J., Rutledge, R. B., Moutoussis, M., and Dolan, R. J. (2017). Neural and computational processes underlying dynamic changes in self-esteem. *eLife*, *6*, e28098. https://doi.org/10.7554/eLife.28098

Wood, J. V., Heimpel, S. A., Newby-Clark, I. R., and Ross, M. (2005). Snatching defeat from the jaws of victory: self-esteem differences in the experience and anticipation of success. *Journal of Personality and Social Psychology*, *89*(5), 764–780. https://doi.org/10.1037/0022-3514.89.5.764

Yu, A. J., and Dayan, P. (2005). Uncertainty, Neuromodulation, and Attention. *Neuron*, *46*(4), 681–692. https://doi.org/10.1016/j.neuron.2005.04.026

# Chapter 6

# General Discussion

# SUMMARY OF EMPIRICAL FINDINGS, LIMITATIONS AND FUTURE DIRECTIONS

## Chapter 2: how inequality affects human motivation to pursue rewards

### *Summary*

The first study investigated how decisions to pursue rewards are altered by a person's awareness that some people in their group were luckier than others in the rewards that they were offered for performing the same task – a situation that reflects opportunity gaps between individuals in everyday life. We found that an unfair distribution of rewards between group members had a negative impact on the decision to pursue rewards of not only disadvantaged individuals but also of advantaged individuals. Separately, we found that offer rank had an independent effect on decisions to pursue rewards. The proposed rank-inequality model outperformed many alternative formulations in explaining participants' reactions to opportunity-gaps. Additionally, we showed the effects of rank and unfairness on willingness to pursue rewards are partially mediated by experienced feelings.

### *Limitations and future directions*

In most real-life situations personal wealth is determined by a mixture of luck and effort, rather than either of these factors alone. Moreover, even when in some situations luck is a more important factor in determining the outcomes, people might still exhibit illusion of control over the outcomes (Presson, and Benassi, 1996), beliefs that the system determining rewards is just (Frank, Wertenbroch, and Maddux, 2015; Smith, 1985), or simply underestimate the role of randomness (Frank, 2016; Taleb, 2007; Teigen and Keren, 2020). These biases could potentially diminish the reactions to unfairness, by creating a false belief of participating in a meritocratic system (Ku, and Salmon, 2013; Frank, 2016). Therefore, future studies will need to determine if the negative effect of unfairness and relative position are also

present in situations where the role of merit and luck is more ambiguous. Such experiment would also help to distinguish the effects of inequality and unfairness that are overlapping in the above study.

## Chapter 3: how do we form perceptions of inequality

### *Summary*

The third study has focused on characterizing computations that govern subjective perceptions of statistical dispersion. We demonstrated that subjective inequality judgments violate several normative axioms used in economics to measure inequality: the anonymity principle, by being affected by their position in the distribution, the scale-independence principle, by being affected by the size of the economy, and the additivity principle, by being insensitive to the addition of incomes that transforms positively skewed distributions into negatively skewed ones. However, we do find that people on average respect the transfer principle, according to which a transfer of income from a richer person to poorer person should decrease inequality. We synthesize these findings by developing a Subjective Equality Index – a parametric model of inequality perception that fits better to participants' data than any other conventional measure of inequality.

### *Limitations and future directions*

The above study focused on inequality judgments. However, it remains to be seen to what extent findings about inequality perception generalize to fairness perception. One study has found that, quite surprisingly, fairness judgments are more consistently aligned with the transfer principle than inequality judgments (Amiel, Cowell, and Gaertner, 2012). No other studies have compared fairness and judgments directly, and with respect to other principles – a gap that will need to be addressed in future research. Despite testing 60 different distributions, we have not tested all possible features that could govern subjective inequality. One example of such omitted aspect is

sensitivity to horizontal translation of the distribution, that is a transformation of the distribution that results from adding a fixed amount of income to all incomes. According to Gini coefficient, such translation should result in decrease in inequality. According to the Subjective Equality Index however, the inequality should not change – a prediction that could be tested in future iterations of the current research.

## Chapter 4: how wealth of foreign countries affects our well-being

### *Summary*

The second study investigated if the human tendency to compare oneself to others extends beyond national borders, and to what degree potential international comparisons could affect how people evaluate their lives. Our results suggest that there is a strong relationship between international relative living standards and life satisfaction, that cannot be explained by absolute living standards or other control variables. Out of 13 different conceptualizations of international comparisons, a rank-inequality fit closest to the well-being data. Additionally, we show that life satisfaction is not only related to relative comparisons with people in person's own country, but also with relative comparisons to people in other countries, comparisons of one's country to other countries, and inequality between countries in a region of the world.

### *Limitations and future directions*

While the effect of country's rank on the international stage on well-being of its citizens is intuitive and consistent with other lines of research, the effect of inequality between countries on well-being is a bit more surprising. One explanation of it could be that inequality between countries in the region makes it more salient for people that living standards depend on lottery of birth, making it harder to believe that a person is living in a just world – a belief that has been shown to serve a protective function for mental health and well-

being (Nartova-Bochaver, Dona, and Ruprich, 2019; Lerner, 1997; Otto et al., 2005). Another possibility is that inequality between countries in the region might be associated with some historical factors, such as international conflicts, that are not captured by present macroeconomic variables but might have had an impact on other unmeasured variables, such as culture. Future studies will need to address these questions by using richer datasets.

## Chapter 5: how do we decide what social feedback are we interested in

### *Summary*

The fourth study investigated the computational mechanisms underlying the information seeking driven by a preference for information that confirms beliefs about oneself – a motive known as self-verification. We find that receiving surprising information about oneself, irrespective if good or bad, induced negative feelings and decreased one's confidence in the specific self-rating. Our results show that a heuristic that relies on a reference-point based on average self-rating fit the closest to participants information-seeking behavior. According to this choice strategy, participants will be more likely to choose positive sources of information the higher they rate their specific personality trait in comparison to how their rate their personality traits on average, and will more likely to choose negative source of information the lower they rate this trait in comparison to how they rate themselves on average. Additionally, we find that the variability between participants in information-seeking strategies could be explained by what strategy is be optimal for a specific participant in minimizing surprises: we show that participants who would benefit most from learning about the sources of information and adjusting their choices accordingly, are also more likely to actually show such behavior.

### *Limitations and future directions*

The above study has found that those whose general self-esteem deviates more from the average received feedback, are more likely to engage in learning strategies – a behavior also suggested by a simulation, in which we have shown that learning strategies are more optimal than heuristics in such situations. This pattern could be either a result of participants switching to learning strategies, once the heuristics prove to be failing to minimize surprises, or alternatively could be a result of inter-individual differences between people with moderate and extreme self-esteem values. Future studies will need to address this question either by developing computational models that allow for dynamic transitions between heuristics and learning, or by a study design that allows for different average feedback value at different times during the experiment. Important limitation of this study is also a fact that participants could assume that each trait evaluation was uncorrelated with past evaluations on other traits – an assumption to some extent justified by Big 5 personality model, in which certain trait groups could be considered orthogonal (Goldberg, 1992). Therefore, the design of the study might have artificially boosted reliance on heuristics, by creating uncertainty about the usefulness of learning.

## SYNTHESIS

### Overview

As stated at the beginning of the current thesis, previous studies have demonstrated that the value of rewards is most likely constructed as a linear combination of the value derived from the properties the offered reward and the value derived from contextual cues (Burke et al., 2016). Despite a plethora of evidence suggesting that context can change how rewards are evaluated, we still know little about the exact mechanisms through which context influences evaluation of rewards. The studies presented in the current thesis

fill in this gap. They do so by testing predictions of many different theories of contextual influence in a social setting, where I was able to quantify the possible influence of the properties of the entire distribution of stimuli on the evaluation of rewards. This is an important methodological advancement, as most previous experimental studies focused on contextual effects in a situation where there are only two stimuli. In such case, all reviewed here theories of contextual influence will predict that a presentation of an additional stimulus with higher intensity (e.g. alternative greater reward), will decrease perceived intensity of the originally presented stimulus. However, the predictions of different theories will start to diverge once there are many presented stimuli at the same time. Our results suggest that when people focus on external context, such as distribution of income in society, they evaluate their own rewards based on their ordinal position in the distribution, and additionally incorporate into this evaluation general preferences about the distribution. On the other hand, when people focus on the internal context, such as beliefs about oneself, they tend to contrast the presented stimuli with the mean. The next section provides a speculative overview of how internal and external context could interact with each other in influencing evaluation of rewards. I then answer some outstanding questions that are of general practical and theoretical interest, and that required a synthesis of evidence presented in this thesis.

## The influence of external and internal social context on valuation of rewards

Brief look at the possible relationships between internal and external social context reveals a complex picture (**Figure 1.**). Here I provide a summary of these relationships, highlighting the links illuminated by studies described in the current thesis, and their relationships with past research.

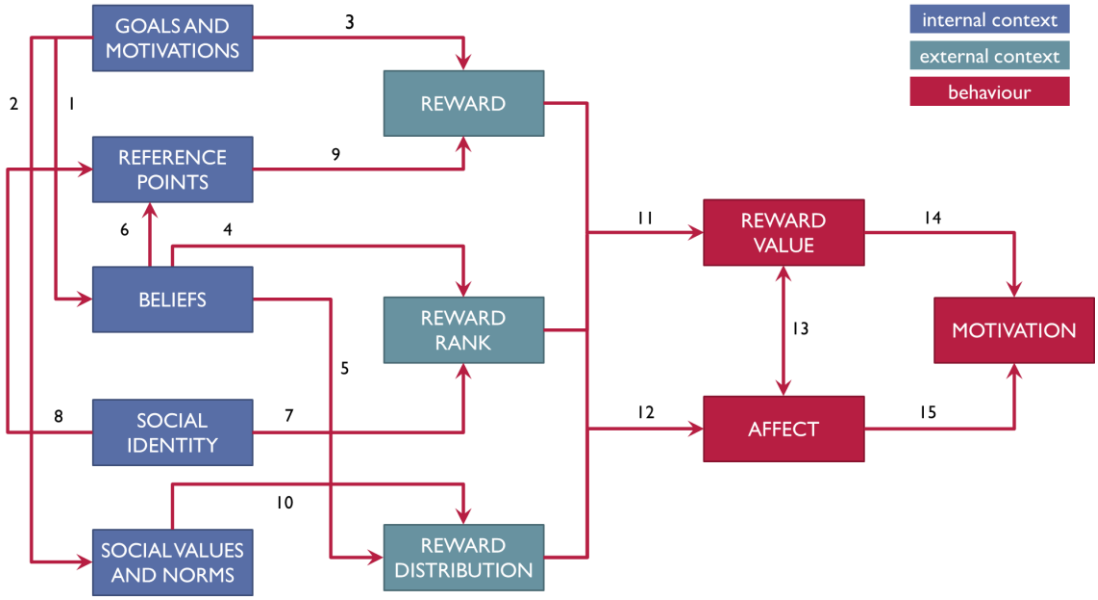


**Figure 1. Overview of relationships between internal social context, external social context and reward value.**

At the most general level it could be argued that internal social context (personal goals, motivations, reference points, beliefs, social identity, and social values) will change the interpretations of the external social context (rewards, their position in the distribution, and the distribution itself), which in turn will affect reward value and its behavioral expression.

The root of these relationships can be traced to personal goals and motivations, which are known to bias the formation of beliefs (1) (Sharot and Garrett, 2016) and preferences for information (Kappes et al., 2020; Charpentier, Bromberg-Martin, and Sharot, 2018; Gesiarz, Cahill, Sharot, 2019). Personal goals and motivations can also influence how people utilize social norms and values (2), often promoting their strategic and opportunistic use in a self-serving manner (Van Baar; Baar, Chang, and Sanfey, 2019; Dana, Weber and Kuang, 2007; Gesiarz and Crockett, 2015). Finally, fulfillment of current goals and motivations will ultimately determine the value of the reward itself (3). As demonstrated in Chapter 5 this sometimes can lead to negative response to positive stimuli, for example when a positive stimulus disconfirms one's beliefs.

Links (4) and (5) emphasize the fact that people react to their subjective beliefs about their rank in the society and the distribution of rewards in society, rather than objective reality. Both over and underestimations of inequality in society and personal position within it has been demonstrated in the literature (Hauser and Norton, 2017; Gimpelson and Treisman, 2017). In Chapter 3 we show how subjective perceptions of reward distributions are to some extent based on objective information, but deviate from normative views present in economics, and are biased by personal position in the distribution.

Personal beliefs, together with social identity, can also influence the reference points and standards that we compare ourselves to (Van Praag, 2010). As shown in Chapter 5, these reference points can change what is considered to be positive or negative. As shown in chapter 4, social identity can also determine the unit of comparison: either individual or group level.

Finally, beliefs will also moderate the influence of social norms and values on evaluations of reward distributions. For example, although a person might be averse to perceived unfairness, they might believe that in general the world is a just place, where most people get what they deserve (Nartova-Bochaver, Dona, and Ruprich, 2019; Lerner, 1997; Otto et al., 2005).

Reward magnitude, its ordinal position in the distribution, and the statistical dispersion of rewards can also change the perceived value of rewards, which will be reflected in both the experienced affect and the motivation to pursue those rewards, as shown in Chapter 2 and 4. Chapter 2 additionally demonstrated that reward value will affect motivation both directly, and indirectly through its' impact on experienced feelings.

**Are people averse to inequality?**

As outlined in the introduction to the current thesis, inequality-aversion hypothesis remains controversial on many grounds. Chapter 2 and 4 contribute to this discussion, strengthening the case for inequality-aversion, or

at least unfairness-aversion. In Chapter 2 we describe an experiment that is free of many factors that were considered to be confounds promoting inequality-averse behavior, without a need for inequality-averse preferences. In this experiment, participants do not engage in re-distribution decisions, or affect the situation of other participants in any other way, eliminating the influence of framing (List, 2007; Bardsley, 2008), reputation concerns (Engelmann and Fischbacher, 2009), reciprocity concerns (Kube, Maréchal, and Puppe, 2012), guilt-aversion (Battigalli and Dufwenberg, 2007), or retribution motives (Suleiman, 1996). Instead participants report their currently experienced feelings and willingness to pursue reward – both of which are negatively affected by inequality, suggesting that it has aversive properties even when the re-distribution behavior is not an option. Important limitation of this interpretation is related to the fact that the inequality in the study was transparently unfair – making it difficult to distinguish the observed effect from unfairness aversion.

The case for inequality/unfairness aversion is further strengthened in Chapter 4, in which we show that that people's reported well-being is negatively affected by both between countries and within country inequality in living standards. Nevertheless, using the same dataset as in Chapter 4, we replicate previously reported inverse U-shape relationship between country's inequality and average well-being, as well as differential effect of inequality in developing and developed countries (**Figure 2.**) (Kelley, and Evans, 2017; Katic and Ingram, 2017), casting doubt on universality of inequality aversion, and pointing to possible moderating factors, such as perceived opportunities stemming from inequality. Together, these findings are more consistent with unfairness-aversion, rather than inequality aversion hypothesis.

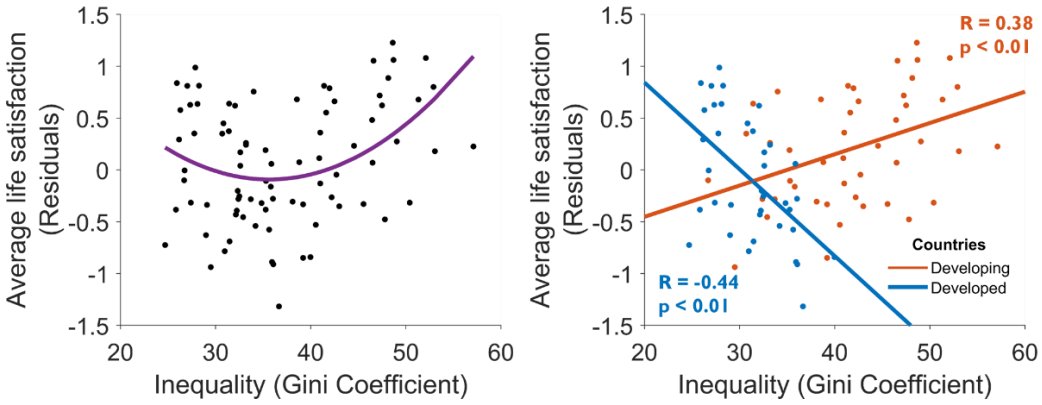


**Figure 2. Relationship between average well-being and inequality within a country.**

### Could influence of inequality be explained by normalization effects?

One intriguing possibility that could underlie many of the observed effects across studies reported in the current thesis, is that it is not the social preferences and inequality aversion that cause devaluation of rewards, but a normalization processes associated with presence of high magnitude rewards in the context, which could be considered to be a perceptual mechanism To test if this is indeed the case, we compared the Rank-inequality model with two most popular value normalization models: divisive-normalization (computed as income divided by sum of incomes) and range-normalization (computed as income divided by the range of incomes) (Soltani, De Martino and Camerer, 2012; Louie et al., 2013). We find that Rank-Inequality model is a better model in both explaining willingness to pursue rewards in data from Chapter 2 and effect of international comparisons on well-being in data from Chapter 4 (**Table 1.**). Although this does not exclude a possibility that inequality could affect evaluation of rewards through some normalization process, it can be concluded that the two dominant models of value-normalization cannot explain the observed patterns of behavior better than the rank-inequality model.

**Table 1. Comparison of rank-inequality model with value-normalization models.**

| Model | BIC Experimental Data | AIC Experimental Data | BIC Gallup Survey | AIC Gallup Survey |
|---|---|---|---|---|
| **Rank-inequality** | 3661.9 | 3596 | 3337 | 3310.1 |
| **Divisive normalization** | 3661.6 | 3608.9 | 3491.8 | 3740.3 |
| **Range normalization** | 3681.7 | 3681.7 | 3490.2 | 3468.7 |

## Overall conclusions

Taken together, the findings presented in this thesis provide a more complete understanding of the interplay between rewards and internal/external social contexts. They indicate that: (i) presentation of rewards in the context of rewards received by others changes the hedonic experience with them, as well as their motivational value; (ii) influence of the social context on rewards is best expressed as a combination of reward rank and inequality between rewards; (iii) perceptions of reward distributions often deviate from normative accounts and personal rank can bias how we perceive such distributions; (iv) social identity and beliefs about oneself can change how people engage in social comparisons, and respond to social feedback.

Throughout the thesis, I attempted to supplement the descriptive results with computational models, and compare predictions of several different theories, providing a first step towards a more mechanistic understanding of the influence of social context on reactions to rewards.

## REFERENCES


Amiel, Y., Cowell, F., & Gaertner, W. (2012). Distributional orderings: An approach with seven flavors. *Theory and Decision*, *73*(3), 381–399. https://doi.org/10.1007/s11238-011-9243-x

Baar, J. M. van, Chang, L. J., & Sanfey, A. G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature Communications*, *10*(1), 1–14. https://doi.org/10.1038/s41467-019-09161-6

Bardsley, N. (2008). Dictator game giving: Altruism or artefact? *Experimental Economics*, *11*(2), 122–133. https://doi.org/10.1007/s10683-007-9172-2

Battigalli, P., & Dufwenberg, M. (2007). Guilt in Games. *The American Economic Review*, *97*(2), 170–176. JSTOR.

Charpentier, C. J., Bromberg-Martin, E. S., & Sharot, T. (2018). Valuation of knowledge and ignorance in mesolimbic reward circuitry. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(31), E7255–E7264. https://doi.org/10.1073/pnas.1800547115

Dana, J., Weber, R. A., & Kuang, J. X. (2007). Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, *33*(1), 67–80. https://doi.org/10.1007/s00199-006-0153-z

Eldar, E., & Niv, Y. (2015). Interaction between emotional state and learning underlies mood instability. *Nature Communications*, *6*(1), 1–10. https://doi.org/10.1038/ncomms7149

Engelmann, D., & Fischbacher, U. (2009). Indirect reciprocity and strategic reputation building in an experimental helping game. *Games and Economic Behavior*, *67*(2), 399–407. https://doi.org/10.1016/j.geb.2008.12.006

Frank, D. H., Wertenbroch, K., & Maddux, W. W. (2015). Performance pay or redistribution? Cultural differences in just-world beliefs and preferences for wage inequality. *Organizational Behavior and Human Decision Processes*, *130*, 160–170. https://doi.org/10.1016/j.obhdp.2015.04.001

Frank, R. H. (2016). *Success and Luck: Good Fortune and the Myth of Meritocracy*. Princeton University Press.

Gesiarz, F., Cahill, D., & Sharot, T. (2019). Evidence accumulation is biased by motivation: A computational account. *PLOS Computational Biology*, *15*(6), e1007089. https://doi.org/10.1371/journal.pcbi.1007089

Gęsiarz, F., & Crockett, M. J. (2015). Goal-directed, habitual and Pavlovian prosocial behavior. *Frontiers in Behavioral Neuroscience*, *9*. https://doi.org/10.3389/fnbeh.2015.00135

Gimpelson, V., & Treisman, D. (2018). Misperceiving inequality. *Economics & Politics*, *30*(1), 27–54. https://doi.org/10.1111/ecpo.12103

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1), 26–42. https://doi.org/10.1037/1040-3590.4.1.26

Hauser, O. P., & Norton, M. I. (2017). (Mis)perceptions of inequality. *Current Opinion in Psychology*, *18*, 21–25. https://doi.org/10.1016/j.copsyc.2017.07.024

Kappes, A., Harvey, A. H., Lohrenz, T., Montague, P. R., & Sharot, T. (2020). Confirmation bias in the utilization of others' opinion strength. *Nature Neuroscience*, *23*(1), 130–137. https://doi.org/10.1038/s41593-019-0549-2

Katic, I., & Ingram, P. (2018). Income Inequality and Subjective Well-Being: Toward an Understanding of the Relationship and Its Mechanisms. *Business & Society*, *57*(6), 1010–1044. https://doi.org/10.1177/0007650317701226

Kelley, J., & Evans, M. D. R. (2017). Societal Inequality and individual subjective well-being: Results from 68 societies and over 200,000 individuals, 1981-2008. *Social Science Research*, *62*, 1–23. https://doi.org/10.1016/j.ssresearch.2016.04.020

Ku, H., & Salmon, T. C. (2013). Procedural fairness and the tolerance for income inequality. *European Economic Review*, *64*, 111–128. https://doi.org/10.1016/j.euroecorev.2013.09.001

Kube, S., Maréchal, M. A., & Puppe, C. (2012). The Currency of Reciprocity: Gift Exchange in the Workplace. *American Economic Review*, *102*(4), 1644–1662. https://doi.org/10.1257/aer.102.4.1644

Lerner, M. J. (1997). What Does the Belief in a Just World Protect Us from: The Dread of Death or the Fear of Undeserved Suffering? *Psychological Inquiry*, *8*(1), 29–32. JSTOR.

List, J. A. (2007). On the Interpretation of Giving in Dictator Games. *Journal of Political Economy*, *115*(3), 482–493. JSTOR. https://doi.org/10.1086/519249

Nartova-Bochaver, S., Donat, M., & Rüprich, C. (2019). Subjective Well-Being From a Just-World Perspective: A Multi-Dimensional Approach in a Student Sample. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01739

Otto, K., Boos, A., Dalbert, C., Schöps, D., & Hoyer, J. (2006). Posttraumatic symptoms, depression, and anxiety of flood victims: The impact of the belief in a just world. *Personality and Individual Differences*, *40*(5), 1075–1084. https://doi.org/10.1016/j.paid.2005.11.010

Presson, P. K., & Benassi, V. A. (1996). Illusion of control: A meta-analytic review. *Journal of Social Behavior & Personality*, *11*(3), 493–510.

Smith, K. B. (1985). Seeing justice in poverty: The belief in a just world and ideas about inequalities. *Sociological Spectrum*, *5*(1–2), 17–29. https://doi.org/10.1080/02732173.1985.9981739

Soltani, A., Martino, B. D., & Camerer, C. (2012). A Range-Normalization Model of Context-Dependent Choice: A New Model and Evidence. *PLOS Computational Biology*, *8*(7), e1002607. https://doi.org/10.1371/journal.pcbi.1002607

Suleiman, R. (1996). Expectations and fairness in a modified Ultimatum game. *Journal of Economic Psychology*, *17*(5), 531–554. https://doi.org/10.1016/S0167-4870(96)00029-3

Taleb, N. N. (2007). *Fooled by Randomness: The Hidden Role of Chance in Life and in the Markets*. Penguin UK.

Teigen, K. H., & Keren, G. (2020). Are random events perceived as rare? On the relationship between perceived randomness and outcome probability. *Memory & Cognition*. https://doi.org/10.3758/s13421-019-01011-6

Van Praag, B. (2011). Well-being inequality and reference groups: An agenda for new research. *The Journal of Economic Inequality*, *9*(1), 111–127. https://doi.org/10.1007/s10888-010-9127-2