Improving the Reliability and Generalizability of Scientific Research

by

Leonid Tiokhin

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2018 by the
Graduate Supervisory Committee:

Daniel J. Hruschka, Co-Chair
Thomas J.H. Morgan, Co-Chair
Robert Boyd
Willem Frankenhuis
Carl Bergstrom

ARIZONA STATE UNIVERSITY

December 2018

ABSTRACT

Science is a formalized method for acquiring information about the world. In recent years, the ability of science to do so has been scrutinized. Attempts to reproduce findings in diverse fields demonstrate that many results are unreliable and do not generalize across contexts. In response to these concerns, many proposals for reform have emerged. Although promising, such reforms have not addressed all aspects of scientific practice. In the social sciences, two such aspects are the diversity of study participants and incentive structures. Most efforts to improve scientific practice focus on replicability, but sidestep issues of generalizability. And while researchers have speculated about the effects of incentive structures, there is little systematic study of these hypotheses. This dissertation takes one step towards filling these gaps. Chapter 1 presents a cross-cultural study of social discounting – the purportedly fundamental human tendency to sacrifice more for socially-close individuals – conducted among three diverse populations (U.S., rural Indonesia, rural Bangladesh). This study finds no independent effect of social distance on generosity among Indonesian and Bangladeshi participants, providing evidence against the hypothesis that social discounting is universal. It also illustrates the importance of studying diverse human populations for developing generalizable theories of human nature. Chapter 2 presents a laboratory experiment with undergraduates to test the effect of incentive structures on research accuracy, in an instantiation of the scientific process where the key decision is how much data to collect before submitting one's findings. The results demonstrate that rewarding novel findings causes respondents to make guesses with less information, thereby reducing their accuracy. Chapter 3 presents

an evolutionary agent-based model that tests the effect of competition for novel findings on the sample size of studies that researchers conduct. This model demonstrates that competition for novelty causes the cultural evolution of research with smaller sample sizes and lower statistical power. However, increasing the startup costs to conducting single studies can reduce the negative effects of competition, as can rewarding publication of secondary findings. These combined chapters provide evidence that aspects of current scientific practice may be detrimental to the reliability and generalizability of research and point to potential solutions.

DEDICATION

To a future where the pursuit of truth is placed above all else.

ACKNOWLEDGMENTS

Thank you to all whose conversations, comments, and support helped to make this dissertation possible. There are too many of you to name here. For starters, I thank my committee members, Dan Hruschka, Tom Morgan, Rob Boyd, Willem Frankenhuis and Carl Bergstrom, for their support, guidance, career advice, and intellectual investment. Special thanks to my advisors Dan Hruschka and Tom Morgan for their continued support in countless ways, and to Rob Boyd, who has motivated me to be a better scholar for my entire academic career. To Willem Frankenhuis, who has invested more in my development than can be reasonably expected of any committee member, who has inspired me over the last decade, and who continues to push me to focus on the 10% not covered by Sturgeon's law: thank you for your continued support and friendship. To Daniel Fessler, thank you for years of friendship, guidance, and investment in my intellectual development. To longtime friends and those who have come and gone, and to family members who have always been there, I can't thank you enough. To all those in Java and West Sumatra who helped me throughout my language studies and fieldwork, thank you for your hospitality and for making my time in Indonesia so memorable. To the open science movement: thank you for inspiring a generation of scholars to try and improve academic science instead of abandoning ship. And to Natalia: Pipos.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

INTRODUCTION

One goal of science is to accumulate information in order to produce increasingly accurate theories of the world. Recently, science's ability to do so has come under scrutiny. In 2005, John Ioannidis published a paper titled "Why most published research findings are false", in which he built a simple analytical model to determine the effects of various factors on the probability that a positive finding corresponded to a true effect (i.e. Positive Predictive Value) (Ioannidis, 2005). Ioannidis found that, for parameter combinations that approximate those in various fields, Positive Predictive Value is less than 50%. In other words, Ioannidis's model concluded that more than half of published positive results are false positives. This work received widespread attention (Aschw, 2015) and for good reason: to the extent that this model captures the core features of scientific practice, we should expect the validity of many scientific findings to be questionable.

One way to determine the validity of research findings is to conduct direct replications. That is, recreate the essential elements of a research study, straying as little as possible from the original design, and determine the ability of the same method to generate the same results upon repetition (Zwaan, Etz, Lucas, & Donnellan, 2018). In the last several years, scholars across the social and biological sciences, in fields including psychology, experimental economics, experimental philosophy, and cancer biology have conducted large-scale replication attempts of published findings. These efforts have made one thing clear: all fields so-far studied have some subset of published findings that cannot be replicated (Begley & Ioannidis, 2015; Camerer et al., 2016; Collaboration &

others, 2015; Cova et al., 2018; Nosek & Errington, 2017; Prinz, Schlange, & Asadullah, 2011).

Given these findings, it is no surprise that there are widespread concerns about the validity of published scientific findings. According to a recent *Nature* survey, most scholars in most fields believe that science is facing a replication crisis (Baker, 2016). Current scientific practice is certainly generating information about the world, but there appears to be much room for improvement.

We have reason to suspect that many failures to replicate published findings are due to the fact that published findings are unreliable (e.g. false-positives). Many current practices that scientists engage in increase the probability of unreliable findings. Studies often lack statistical power (Button et al., 2013; Cohen, 1962; Ioannidis, Stanley, & Doucouliagos, 2017; Smaldino & McElreath, 2016) meaning that they have a low probability of accurately detecting a true effect when one exists. And one consequence of low statistical power is, conditional on publication bias against negative results, published research will have a higher ratio of false-positive-to-true-positive findings (Ioannidis, 2005), and produce inflated published estimates of true effect sizes (Button et al., 2013). Second, researchers introduce many sources of undisclosed bias during the research process (i.e. researcher degrees of freedom) that increase the probability of false-positive results (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011). And in part due to publication bias against negative results and lack of incentives for conducting replications, most published studies report novel, positive results (Fanelli, 2011; Makel, Plucker, & Hegarty, 2012).

However, published findings can also fail to replicate even if the original findings were correct because of differences between original studies and replications (Gilbert, King, Pettigrew, & Wilson, 2016). For instance, seemingly arbitrary aspects of experimental context (e.g. light vs. dark room) or experimental design (e.g. using different scales to measure a psychological construct) can generate different results, essentially serving as "hidden moderators" of effects (Frey, Pedroni, Mata, Rieskamp, & Hertwig, 2017; Gilbert et al., 2016; Landy & others, n.d.; N. Schwarz & Clore, 2016; Van Bavel, Mende-Siedlecki, Brady, & Reinero, 2016). Additionally, because most social-science research relies on convenience samples (i.e. college undergraduates) and is otherwise conducted with participants from a narrow-range of humanity (i.e. those from western, educated, industrialized, rich, and democratic societies), even findings that reliably replicate among these participants may not generalize to humanity as a whole (Henrich, Heine, & Norenzayan, 2010; Medin, 2017; Nielsen, Haun, Kärtner, & Legare, 2017). These combined considerations suggest that research findings will often be limited in their reliability and generalizability, and that current scientific practices may exacerbate these problems.

This dissertation research strives to reveal limitations of current research practices and understand their causes. Chapter 1 presents a cross-cultural study of social discounting - the purportedly fundamental human tendency to sacrifice more for socially-close individuals. This study was conducted among the most diverse populations to date (U.S. undergraduates, rural Indonesians, rural Bangladeshis). This allowed examination of whether social discounting generalizes beyond the narrow range of participants used in

previous studies (i.e. college-undergraduates and U.S. participants). In contrast to most prior research, this study found no independent effect of social distance on generosity among Indonesian and Bangladeshi participants, despite documenting this effect among U.S. participants. This finding suggests that social discounting is less generalizable than previously assumed. It also suggests the importance of increasing investment in strong checks on generalizability across diverse human populations, in addition to current scientific reforms that largely focus on the reliability of published findings.

Chapters 2 and 3 move from empirical demonstrations of issues with current scientific practice towards understanding the causes of problematic practices. Chapter 2 presents a laboratory experiment with U.S. undergraduates testing how incentive structures can affect research quality. This experimental protocol was reviewed and provisionally accepted for publication in *Royal Society Open Science* via a new article format called a "registered report" (C. D. Chambers, 2013; Nosek & Lakens, 2014). Registered reports are evaluated based on the importance of the research question and the quality of the proposed methodology, before data-collection ensues. In this way, registered reports ameliorate some problems with how research is currently conducted and published (e.g. incentives to hunt for positive and clean results; bias against null-results; underpowered studies) (Nosek & Lakens, 2014). The experiment in Chapter 2 focuses on the effect of one incentive in particular: rewards for novel results. It tests whether rewarding novel results affects the amount of information that people acquire when solving research questions and how much effort they invest in doing so. This study finds that rewarding novel research findings has harmful effects (i.e. individuals make

guesses with less information) and finds no evidence of benefits (i.e. competition does not cause individuals to increase effort in order to acquire information more efficiently).

Chapter 3 presents an agent-based model that tests the effect of competition for novel research findings on the sample size of studies that researchers conduct. In this model, scientists investigate a phenomenon and compete to be first to obtain a statistically significant result. Scientists can increase statistical power by using larger samples, but this takes more time and so increases their risk of being "scooped". The model demonstrates that competition for novel research findings causes the cultural evolution of research with smaller sample sizes and lower statistical power than when competition is absent. It also finds that increasing the time costs associated with setting up a single study can reduce the negative effects of competition, as can rewarding publication of non-novel findings. Chapters 2 and 3 suggest that current hypotheses about the detrimental effects of competition on the scientific process are valid under some conditions: competition for novel findings can cause reduced research quality. However, competition can also have benefits, causing individuals to increase research quality. Additionally, the model in Chapter 3 points to one way to ameliorate the negative effects of competition: increase rewards for publication of non-novel results and increase the startup costs to conducting studies. These provide a way forward as the scientific community works towards a better understanding of how to change current incentive structures in order to increase the reliability of published findings.

CHAPTER 1

GENERALIZABILITY IS NOT OPTIONAL: INSIGHTS FROM A CROSS-CULTURAL STUDY OF SOCIAL DISCOUNTING

Abstract

Current scientific reforms focus more on solutions to the problem of reliability (e.g. direct replications) than generalizability. Here, we use a cross-cultural study of social discounting to illustrate the utility of a complementary focus on generalizability across diverse human populations. Social discounting is the tendency to sacrifice more for socially-close individuals—a phenomenon replicated across countries and laboratories. Yet, when adapting a typical protocol to low-literacy, resource-scarce settings in Bangladesh and Indonesia, we find no independent effect of social distance on generosity, despite still documenting this effect among U.S. participants. Several reliability and validity checks suggest that methodological issues alone cannot explain this finding. These results illustrate why we must complement replication efforts with investment in strong checks on generalizability. By failing to do so, we risk developing theories of human nature that reliably explain behavior among only a thin slice of humanity.

Introduction

A long-term goal of psychological science is to produce robust and generalizable theories of human nature (Crandall & Sherman, 2016; Rozin, 2009). In recent years, we have become increasingly aware of how inefficiencies in the scientific process obstruct progress towards this goal (Begley & Ioannidis, 2015; C. Chambers, 2017; Ioannidis,

2005; Munafò et al., 2017) and may contribute to the unreliability of published research findings (Begley & Ioannidis, 2015; Camerer et al., 2016; Collaboration & others, 2015). This realization has inspired a revolution: Psychologists (along with scholars from diverse disciplines) have united to propose a range of innovative reforms to current scientific practice (C. Chambers, 2017; Munafò et al., 2017). One central reform has been an increased emphasis on direct replications, studies that recreate the essential elements of previous research to assess the ability of a specific method to generate the same results upon repetition (Zwaan et al., 2018). Direct-replication efforts are rapidly changing the scientific landscape. Large-scale interdisciplinary replication teams are emerging across the globe, scientific journals are increasingly publishing direct replications, and federal governments are beginning to invest resources in replication efforts (Baker, n.d.; Nelson, Simmons, & Simonsohn, 2017).

This cultural shift is long overdue: direct replications are essential for determining the reliability of findings (Koole & Lakens, 2012; Zwaan et al., 2018). Nonetheless, direct replications are no panacea (Munafò & Smith, 2018; Stroebe & Strack, 2014): they often do not reveal boundary conditions for an effect (Fiedler, 2011; Simons, Shoda, & Lindsay, 2017a), the extent to which it replicates under different operationalizations of theoretical constructs (Crandall & Sherman, 2016), or how well it generalizes to different study populations (Henrich et al., 2010). Recently, Marcus Munafo and George Davey Smith (scholars who have dedicated their careers towards increasing scientific reliability) expressed concerns that an increased emphasis on direct replication is "...laudable, but insufficient" (14, p.1) because it underemphasizes the fact that "strong theories emerge

from the synthesis of multiple lines of evidence" (14, p. 3). In other words, direct replications solve one barrier to developing broadly-relevant theories of human nature (i.e. reliability) but do not (and are not designed to) solve others (e.g. generalizability).

Moving generalizability into the limelight

The vast majority of proposals to improve scientific practice sidestep the issue of generalizability. Proposals by leading scholars in the field, including John Ioannidas's "How to Make More Published Research True" and an interdisciplinary "Manifesto for Reproducible Science" have largely focused on threats to reproducibility (e.g. p-hacking, publication bias, low-statistical power, openness of materials, replication) (Ioannidis, 2014; Munafò et al., 2017). Those proposals that have directly engaged with concerns about generalizability have focused largely on experimental design and statistical analysis. For instance, radical randomization of experimental parameters (Baribault et al., 2018) and crowdsourcing operationalizations of theoretical constricts (Landy & others, n.d.) and analytical choices (Silberzahn et al., 2017) have all been proposed as ways to reveal how effects vary due to arbitrary choices that researchers make when designing studies. Here we focus on another longstanding proposal to improve generalizability: increasing sample diversity (Henrich et al., 2010; Moshontz et al., 2018).

Social scientists have long worried that convenience samples (e.g. college undergraduates) bias our inferences about human nature (Arnett, 2008; Henrich et al., 2010; McNemar, 1946; Nielsen et al., 2017; Peterson, 2001; Sears, 1986). Despite these concerns, most social-science research continues to rely on participants from a narrow slice of humanity (i.e. those from western, educated, industrialized, rich, and democratic

societies (Henrich et al., 2010)) and leading proposals for scientific reform scarcely mention the issue of sample diversity (Munafò et al., 2017, but see Moshontz et al., 2018; Simons et al., 2017a). It is not clear why worries about unrepresentative participants have not translated into tangible changes to scientific practice, whereas worries about unreliable effects have: Henrich, Heine, and Norenzayan (2010) published their widely-cited "WEIRD" paper (Henrich et al., 2010) just one year before Simmons, Nelson, and Simonsohn published "False-Positive Psychology" (Simmons et al., 2011), Daryl Bem's published his infamous pre-cognition paper (Bem, 2011), and Diedrik Stapel admitted to fabricating decades worth of psychological data (Bhattacharjee, 2013). The latter three events "drove psychological science into a spiral of methodological introspection" (10, p. 3). The former seems to have led largely to brief caveats that acknowledge the unrepresentativeness of participants, cite the WEIRD paper, and go about business as usual (Medin, 2017; Nielsen et al., 2017).

In this paper, we use our recent multi-site investigation of social discounting as one in an emerging set of case studies to illustrate how failing to conduct checks on generalizability across diverse samples can lead to the production of narrow models of human behavior (for other examples, see (Apicella & Barrett, 2016; Henrich, 2015; Henrich et al., 2010)). We do so acknowledging the fact that convenience samples (including WEIRD populations) are often useful: all authors of this paper have and continue to rely on convenience samples in our work. We also do not have a special reason for choosing social discounting as a case study, besides the fact that we have conducted social-discounting research and are familiar with the literature. Rather, we

suspect that social discounting research illustrates an issue that will become increasingly relevant to many fields of psychological science: we have good evidence that a phenomenon reliably replicates in a limited set of conditions but know little about whether it constitutes a general feature of human nature or is a quirk of the WEIRD participants upon which we so heavily rely.

Social Discounting

In psychology, social discounting is defined as the tendency to bear greater costs to benefit socially-close individuals than socially distant ones (Jones & Rachlin, 2006). Specifically, when given the option to sacrifice money (or some other resource) in order to provide money (or some other resource) to others, people sacrifice substantially more for socially-close partners. Over 50 published studies in the last 10 years document this behavioral bias (see https://osf.io/cfkdr/), including a recent pre-registered direct replication (see https://osf.io/fn9am/). The apparent regularity of a hyperbolic relationship between social distance and generosity has led scholars to hypothesize a fundamental relationship with time discounting (Jones & Rachlin, 2006) and investigate its neural basis (Strombach et al., 2015). Others give social-discounting law-like status (Goeree, McConnell, Mitchell, Tromp, & Yariv, 2010), or describe it as a "robust phenomenon, with respondents across settings and cultures reliably willing to sacrifice more resources for socially close others relative to distant others" (36, p. 1).

From the perspective of assessing reliability, using the same method to find a recurring hyperbolic relationship between social distance and willingness to sacrifice could be considered a success. Yet, we know little about whether successful replication

extends beyond the limited range of participants and methods in these studies. To understand the scope of this potential problem, we reviewed all social-discounting studies that cited Rachlin and Jones' seminal social-discounting paper (Jones & Rachlin, 2006) and used a comparable protocol (see https://osf.io/k8sbg/). Of 43 groups of participants from 21 publications, 40 groups were from the United States and/or university students, with only 3 groups as exceptions (1 from an Indian Mechanical Turk sample, (Hackman, Danvers, & Hruschka, 2015) and 2 from one study in Singapore (Pornpattananangkul, Chowdhury, Feng, & Yu, 2017)). Even research cited as supporting the cross-cultural reliability of social discounting typically relies on university students (Ishii & Eisen, 2018; Ma, Pei, & Jin, 2015; Strombach et al., 2014)[1].

The Standard Social-Discounting Protocol

The standard protocol for assessing social discounting was developed with U.S. college undergraduates ((Jones & Rachlin, 2006); a similar protocol is used in evolutionary psychology to study welfare-tradeoff ratios (Delton & Robertson, 2016)). Typically, it consists of a paper-and-pencil task where participants imagine a list of 100 people closest to themselves. Participants then identify a person (recipient) at a specific location on that list (e.g. #1, #2, #5, #10, #20, #50, #100). For each recipient, participants

[1] One exception is a study that compared social discounting among U.S. college students, urban Chinese, and Kenyan herders (P. Boyer, Lienard, & Xu, 2012). Boyer, Lienard & Xu found that social distance had a weaker effect on generosity among Kenyan herders compared to the other two groups. However, because this study did not measure social distance directly (instead using culture-specific categories such as "high-school friend" or "same-age-set"), it is unclear how to compare its results to those of typical protocols (or even across languages or cultures) and it did not meet our pre-registered exclusion criteria (see https://osf.io/k8sbg/).

make several decisions about keeping some amount of money for themselves or transferring some amount to the recipient. In a typical task (Locey, Jones, & Rachlin, 2011; Rachlin & Jones, 2008; Vekaria et al., 2017), participants might choose between option A and option B as follows:

"A. $85 for you alone. B. $75 for the #___ person on the list.

A. $75 for you alone. B. $75 for the #___ person on the list.

A. $65 for you alone. B. $75 for the #___ person on the list.

…

A. $5 for you alone. B. $75 for the #___ person on the list."

To assess individual generosity, analyses typically calculate the "crossover point" in the sequence of questions where respondents switch from response A (i.e. the selfish option) to response B (i.e. the generous option). For example, if a participant chooses the selfish option at $85 and $75 but switches to the generous option at $65, her crossover point is $70. This approach assumes that participants will switch from generous to selfish only once in a sequence of decisions. Participants that make multiple crossovers are labelled "inconsistent" and often excluded from analyses (Jones & Rachlin, 2006; Vekaria et al., 2017).

To assess ecological and cultural moderators of social discounting, we adapted this social-discounting task (Rachlin & Jones, 2008) to a low-resource, rural, semi-literate setting in Bangladesh, as well as the most commonly studied population in the literature —U.S. college undergraduates. Doing so revealed a drastically different pattern of responding (Table 1; Figures 1 - 3). Unlike U.S. college undergraduates, Bangladeshi participants were not more generous to socially-close partners. We were surprised by this finding: based dozens of prior studies (see Figure 3) we expected to find at least some

degree of social discounting. We then took advantage of an opportunity to run the same task in another non-western, low-resource setting: rural Indonesia. Again, we found the same pattern: Indonesian participants were not more generous to socially-close partners. These patterns could not be explained by several common methodological critiques, such as participants failing to understand the protocol, poor measurement of dependent or independent variables, floor effects, or low statistical power.

Methods

Because participants in rural Bangladesh and Indonesia are semi-literate and have varying levels of schooling and experience with typical abstracted paper-and-pencil tasks, adapting the standard protocol to these settings required extensive modifications (Hruschka, Munira, Jesmin, Hackman, & Tiokhin, in press). These included 1) translation of materials, piloting, and comprehension checks, 2) identifying locally appropriate idioms for ranking relationships by social distance, 3) limiting the list of socially-close individuals to 20, 4) asking respondents to list and then physically rank cards with 20 socially-close individuals rather than asking for abstract rankings of 1,2,5,10, and 20, 5) modifying how participants identify partners (Bangladesh: choosing among images of all individuals in the village; Indonesia: writing and ranking socially-close individuals on notecards), 6) use of an alternative currency (Bangladesh: rice instead of money) to avoid harming ongoing relationships with community members, 7) presenting choices between selfish and generous options on slips of paper that could be placed in a transparent lottery, 8) visual representation of the stakes on slips of paper and 9) using strategically-placed screens to enhance anonymity of decisions. Such challenges

are the norm when developing protocols that are meaningful across variable contexts and cultural settings (Apicella, Crittenden, & Tobolsky, 2017; Henrich et al., 2010; Hruschka et al., in press).

Unless otherwise stated, the following adapted protocol was used uniformly across all three sites, including the U.S. sample. Participants made a list of the 20 people that they felt closest to and that did not live in their same household. We identified the idiom most closely aligned with "social closeness" in Bangladesh and Indonesia from conversations with local respondents about how they describe good relationships where partners care about and help each other (i.e. Bangladesh: *ghonishto,* meaning "thick/viscous"; Indonesia: *dekat*, meaning "close"). Participants then sorted the listed individuals in order of how close they felt to each one. The experimenter then selected individuals at 5 social distances (#1, #2, #5, #10, #20) for the subsequent task. For each of these 5 individuals, participants made 6 dichotomous choices between keeping a certain amount of currency for themselves (i.e. selfish option) or giving a certain amount of the currency to that recipient (i.e. generous option). The generous option remained fixed for all choices. The selfish option varied between an amount equal to the generous option to an amount 10% of the generous option, and the order in which it was presented was randomized. The maximum generous option was scaled to the equivalent of a half-to-full day's wage in each of the contexts (150 Tk in Bangladesh, 25 USD in college student sample, 50,000 IDR in Indonesia). To assess individual decisions with unfamiliar partners, participants also made decisions between selfish and generous options for an

"unknown person". Participants in Bangladesh also made decisions between selfish and generous options for an "acquaintance" in the village.

Each decision was presented as a choice between two paper tickets, one with the selfish option and one with the generous option. For each choice, participants placed their preferred ticket in a small bucket labeled "lottery", and their non-preferred ticket in a small bucket labeled "trash". Participants were instructed that one of the tickets placed in the "lottery" would be paid out at the end of the experiment, whereas all tickets placed in the "trash" would be thrown away. All decisions were made behind a screen so that the experimenter was blind to participant decisions. Participants were instructed that their choices were anonymous and choice order was randomized between individuals.

We found high rates of "inconsistency" in choices (i.e. multiple crossover points for at least 1 recipient) among participants in both Bangladesh and Indonesia (80% and 100% of participants with non-zero generosity were inconsistent in Bangladesh and Indonesia, respectively; see Appendix). As a consequence, we used a weighted average of respondents' 6 decisions (henceforth "expected sharing"). Expected sharing is monotonically increasing with the crossover point when respondents have a single crossover point, does not force exclusion of inconsistent respondents, and provides a simple measure of individual differences in generosity (See Appendix for details and for re-analysis using approximated crossover points). Notably, "inconsistent" respondents did not behave differently than "consistent" respondents, indicating that inconsistency does not arise from failure to comprehend the task (See Appendix).

Upon finishing the task, participants completed several self-report measures. Participants indicated the relative financial need of the recipient compared to themselves via an ordinal scale (i.e. “This person is needier than you”, “This person has the same need as you”, “This person is less needy than you”). Participants in the U.S. and Indonesia also completed the Inclusion of Other in the Self (IOS) Scale (Aron, Aron, & Smollan, 1992), as a complementary measure of closeness to the recipient. Bangladesh participants had difficulty understanding the IOS. As such, in Bangladesh, we developed a protocol using bins of varying distance from the informant, in which villagers could place photos (Hackman, Munira, Jasmin, & Hruschka, 2017; Hruschka et al., in press). Participants also indicated their age and sex. At the end of the experiment, one choice was randomly selected from the lottery for payout. If a selfish option was selected, participants received that payment immediately, along with their participation fee. If a generous option was selected, participants only received their participation fee. After all participants completed the task, experimenters paid the appropriate amount to any recipient randomly selected to receive a payout without indicating who it came from.

A total of 284 participants across 3 sites (U.S. = 40, Bangladesh = 200, Indonesia = 44) participated in this study (See Appendix for demographic characteristics). In the U.S., we recruited 40 participants via emails sent to a list of 6000 undergraduates, curated by the Center for Behavior, Institutions, and the Environment. In Bangladesh, we recruited one participant from each of 200 households across four villages in Northwestern Bangladesh (see 47). In Indonesia, we recruited 44 participants using opportunity sampling from a single rural settlement (*nagari*) in West Sumatra, near the

city of Payakumbuh in the Lima Puluh Kota regency, limiting recruitment to 2 individuals from the same household.

For the U.S. and Indonesia, sample size was determined based on the sample size in a previous lab study of social discounting (Hackman et al., 2015). For Bangladesh, sample size was chosen to provide sufficient power (Power = 0.80, $\alpha = 0.05$) to detect a bivariate association between social distance and generosity with a coefficient of prediction greater than 0.15.

Analysis and Results

We tested whether generosity declines with increasing social distance by regressing expected sharing on social distance using a multilevel model (Table 1). This model controls for genetic relatedness (linear) and relative need (categorical), and for correlated observations from the same participant with random effects for each individual. It also includes two random slopes: social distance and recipient need (See Appendix for alternative-model comparisons). Because the relationship between money forgone and social distance typically follows a heavy-tailed function, we use the natural log of social distance as a predictor. To adjust for multiple comparisons, we use Bonferroni-adjusted alpha ($\alpha = 0.004$) levels for all tests of statistical significance based on 12 tests.

For comparability with prior social-discounting research using real stakes, we limit our analysis to decisions for social distances #1, #2, #5, #10, and #20 (Locey et al., 2011). In the Appendix, we also report analyses including generosity towards an "unknown person" in all 3 sites (See Appendix).

Consistent with prior research, we find a strong independent effect of social distance on generosity among U.S. undergraduates, after controlling for need and relatedness [β = -0.10, 95% CI [-0.12, -0.08], $p<0.001$]. In contrast, we find no independent effect of social distance on generosity among Bangladeshi [β = 0.00, 95% CI [-0.01, 0.01], p = 0.677] or Indonesian [β = -0.01, 95% CI [-0.03, 0.02], p=0.626] participants (Table 1; Figure 1a). The independent effects of social distance on generosity in Bangladesh [β = 0.10, 95% CI [0.08, 0.12], $p<0.001$] and Indonesia [β = 0.10, 95% CI [0.07, 0.12], $p<0.001$] were significantly different from the effect among U.S. undergraduates (see Table 8S in appendix). The independent effects of social distance on generosity in Bangladesh and Indonesia were not significantly different from each other [β = -0.01, 95% CI [-0.03, 0.02], p=0.541]. The maximum plausible effect-size estimates in Bangladesh [β = -0.01] and Indonesia [β = -0.03] are several times smaller than the minimum plausible effect-size estimate in the U.S. [β = -0.08]. These results are robust to a variety of alternative model specifications (see appendix for sensitivity analyses and BIC-approximated Bayes Factors (Wagenmakers, 2007)).

Figure 2 plots the distribution of expected sharing in all sites, as a function social distance and the relative need of a recipient. This depicts a slight decrease in generosity with increasing social distance among Indonesian participants in the raw data. However, the apparent decrease in generosity with increasing social distance among Indonesian participants is due to a confounding effect of relative need judgments at varying social distances: we find this effect only when removing relative need as a fixed-effect predictor from the multilevel model in Table 1 (See Appendix).

This study's finding—that generosity does not increase with decreasing social distance—diverges from well-established findings in social-discounting research. Figure 3 plots our findings against those of comparable social-discounting studies, comprising 39 groups of participants from 19 published articles (for data and pre-registration, see https://osf.io/k8sbg/). For comparability, we calculate the ratio of the maximum amount forgone by participants to the amount transferred to recipients in all studies.

	U.S. Expected Sharing		Bangladesh Expected Sharing		Indonesia Expected Sharing	
	Estimate (CI)	*P*	*Estimate (CI)*	*P*	*Estimate (CI)*	*P*
Fixed Effects						
(Intercept)	0.71 (0.57, 0.84)	<.001	0.15 (0.07, 0.22)	<.001	0.67 (0.56, 0.78)	<.001
Ln Social Distance	-0.10 (-0.12, -0.08)	<.001	0.00 (-0.01, 0.01)	.677	-0.01 (-0.03, 0.02)	.626
Need						
Recipient Equally Needy	-0.10 (-0.22, 0.02)	.121	-0.07 (-0.15, 0.01)	.084	-0.20 (-0.29, -0.10)	<.001
Recipient Less Needy	-0.19 (-0.32, -0.07)	.004	-0.13 (-0.21, -0.05)	.001	-0.31 (-0.42, -0.20)	<.001
Relatedness	0.05 (-0.08, 0.19)	.459	-0.01 (-0.10, 0.08)	.857	0.12 (-0.00, 0.25)	.056

Table 1| Generosity as a function of social distance, need, and relatedness. Fixed effects from multilevel model of social distance, recipient need, and relatedness regressed on expected sharing. Model controls for correlated observations from the same participant with random effects for each individual

and includes random slopes for social distance and recipient need. CI = 95% confidence intervals.

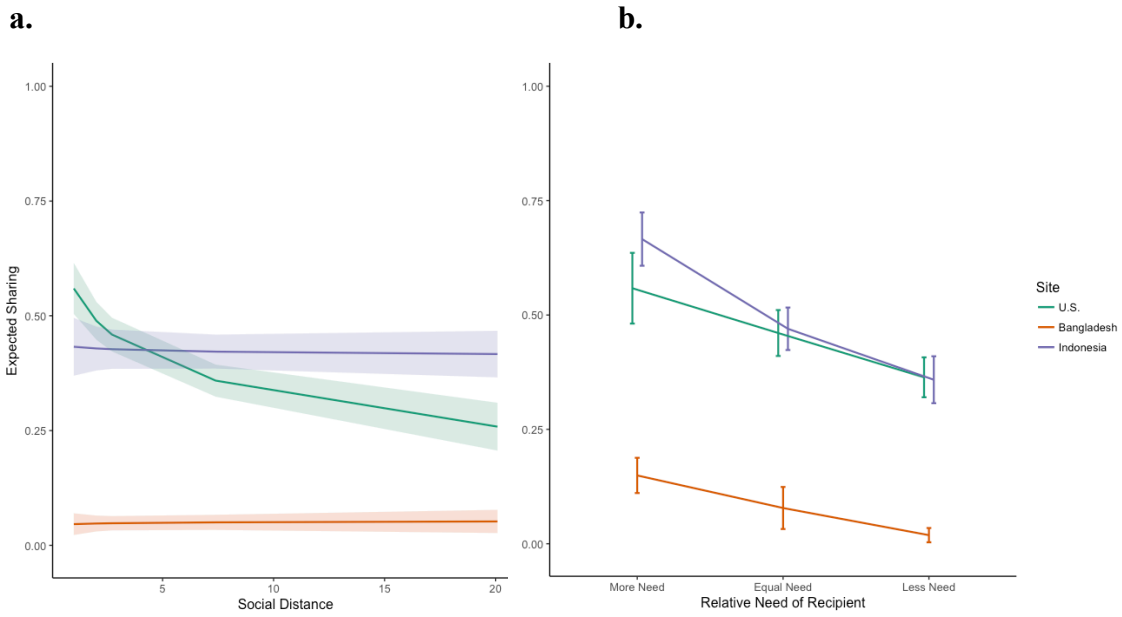
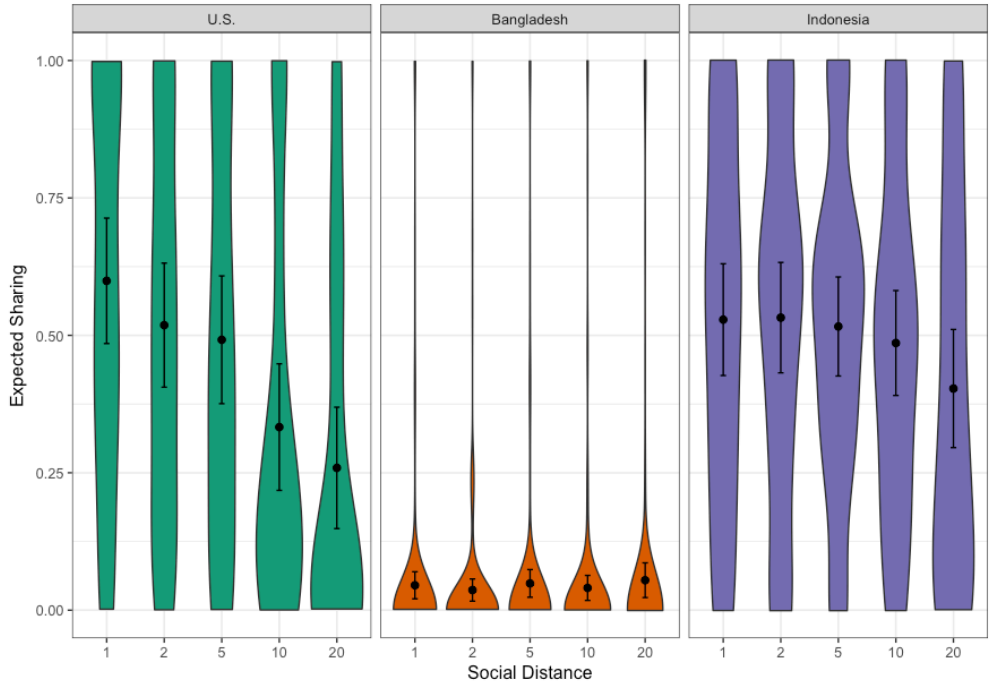


Figure 1| Independent effects of social distance and need on generosity. Independent effects of **a.** social distance (natural log transformed) and **b.** relative need on expected sharing. Model estimates from the multilevel model in Table 1. Error bars represent 95% confidence intervals.

a.



b.

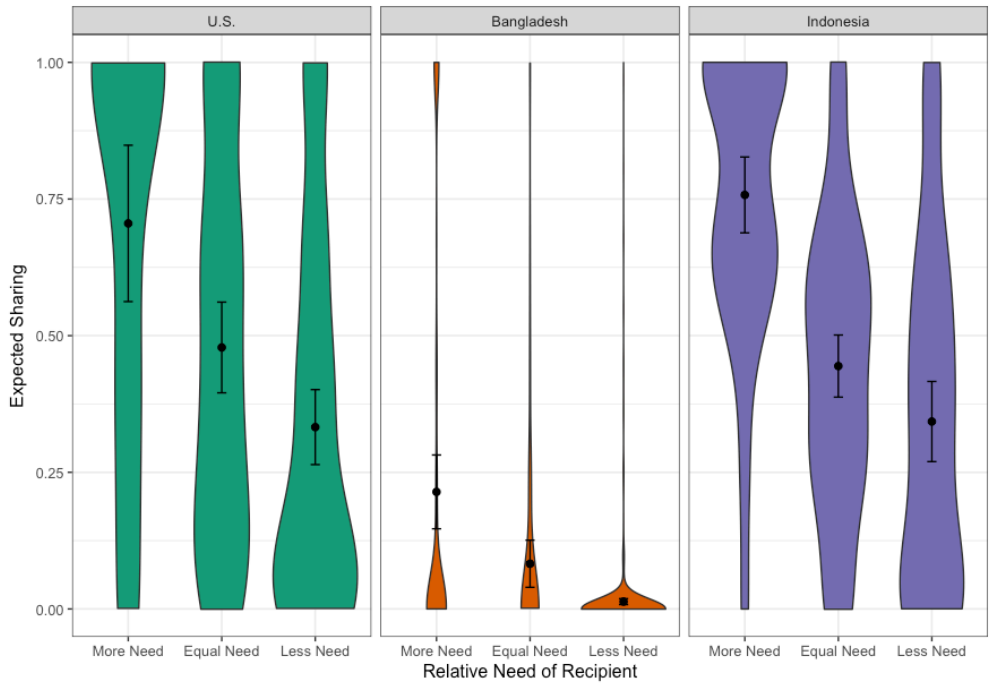


Figure 2| Distribution of generosity as a function of social distance and relative need of recipient. Probability density of expected sharing as a function of **a.** social distance and **b.** relative need. Dots represent arithmetic means. Error bars represent 95% confidence intervals.

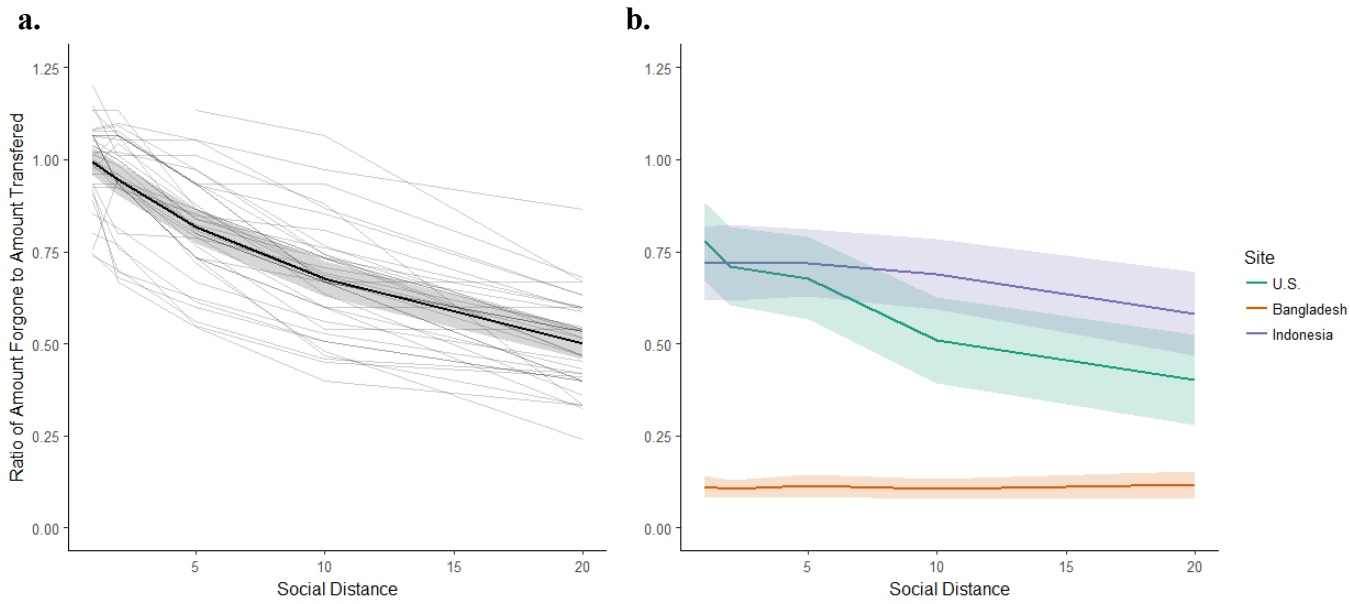


Figure 3| Social discounting in prior research and current study. Social discounting, plotted as the ratio of the maximum amount foregone by participants to the amount transferred to recipients. Error bars represent 95% confidence intervals. **a.** Prior studies (N = 39 groups of subjects across 19 publications). **b.** Current study.

To further probe why social distance did not predict generosity among Bangladeshi and Indonesian participants, we analyzed participants' verbal statements about the reasons for their decisions. In all sites, we asked participants to explain their decisions at the end of the task. Bangladeshi and U.S. participants were asked to justify their decisions towards each recipient, whereas Indonesian participants were asked for justifications once, after making all decisions towards all recipients. After reading a subset of statements, we generated a codebook with 13 categories (See Appendix). Each author independently coded all participant statements in all sites and we resolved conflicting codes via collaborative discussion. We then analyzed the subset of codes most relevant to our findings: statements about qualities of relationships and statements about own or recipient need (see Appendix for codebook and https://osf.io/cfkdr/ for

complete data set). Figure 4 plots the proportion of participants who made at least 1 statement about relationships or need in justifying their decisions. The majority of participants in all sites mentioned need as a justification for their decisions. In contrast, only in the U.S. did the majority of participants mention relationships. In Bangladesh and Indonesia, 15% and 45% of participants mentioned relationships, respectively. In Indonesia, many respondents mentioned both need and social relationships, even though we did not find an effect of social distance on generosity [β = -0.01, 95% CI [-0.03, 0.02], p=0.626]. However, almost all Indonesian participants that mentioned relationships only mentioned the importance of helping family, and Indonesia was the site where we found the largest estimate for the effect of relatedness on generosity [β = 0.12, 95% CI [-0.00, 0.25], p=0.056]. Participants' statements appear to correspond closely to their behavior in the social-discounting task, providing convergent evidence that factors regarding relationship quality have a stronger impact on U.S. participants' behavior than they do for Bangladeshi and Indonesian participants.

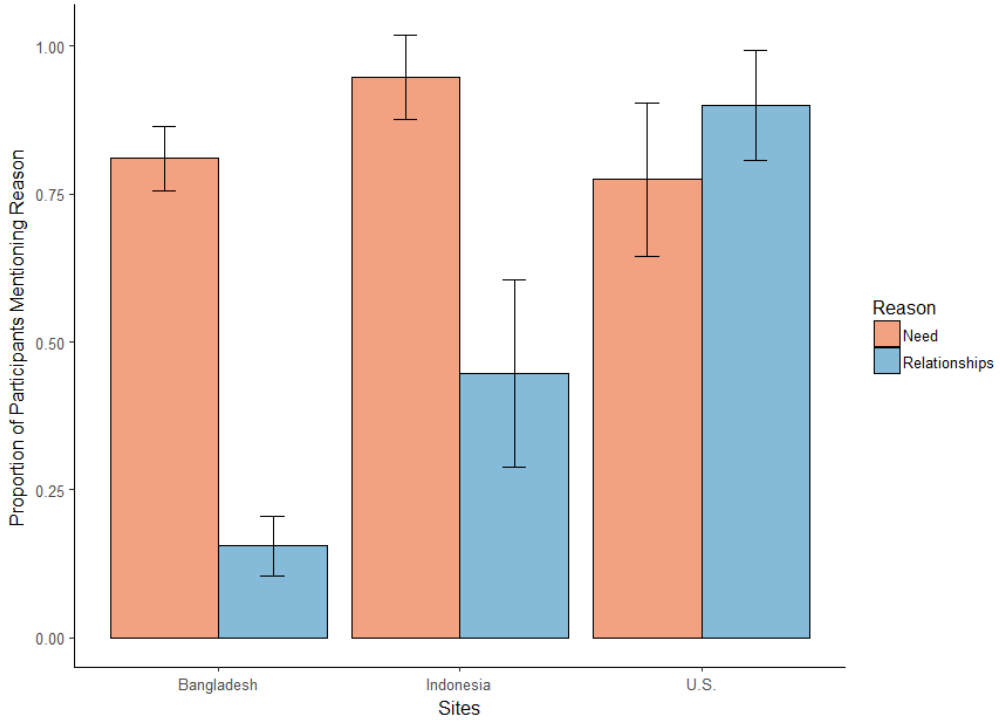


Figure 4| Proportion of Participants Who Mentioned Need or Relationships When Explaining their Behavior. Proportion of participants that explained their behavior in the social-discounting task by making at least 1 statement about qualities of social relationships (Relationships) or own/recipient need (Need).

We found that Bangladeshi participants displayed low levels of generosity (160 of 200 participants did not give anything to anyone). To check whether our findings were affected by the inclusion of these participants, we re-ran the same multilevel model on the subset of participants with non-zero levels of generosity (U.S., n = 39; Bangladesh, n = 35; Indonesia, n = 42). The results were robust to these exclusions (See Appendix).

One potential reason for the lack of a social discounting effect in Bangladesh and Indonesia may be that the dependent variable (i.e. amount foregone) was unreliably measured. However, another theoretically-relevant covariate — recipient's need relative to participant's need— showed significant independent associations with the dependent variable in all 3 sites. Specifically, participants were more generous to individuals

classified as "more needy" versus those classified as "less needy" in all 3 sites (U.S. [β = -0.19, 95% CI [-0.32, -0.07], p = 0.004], Bangladesh [β = -0.13, 95% CI [-0.21, -0.05], p=0.001], Indonesia [β = -0.30, 95% CI [-0.41, -0.20], p < 0.001]). This suggests that generosity was measured with sufficient reliability to have substantial and significant associations with at least one theoretically-important variable.

Another reason for the lack of a social discounting effect may be that the independent variable (i.e. social distance) was unreliably measured, despite having shown strong relationships with reported helping in a previous study in the same Bangladesh context (Hackman et al., 2017). To assess this possibility, we analyzed the relationship between social-distance rankings and reports of closeness via the IOS scale (and Bangladesh modification of the IOS), using a multilevel model with random effects for each individual. Participants in the U.S. (β = -1.23, 95% CI [-1.38, -1.08], p<0.001), Indonesia (β = -0.86, 95% CI [-1.00, -0.72], p<0.001) and Bangladesh (β = -0.78, 95% CI [-1.00, -0.59], p<0.001) reported feeling less close to individuals at greater social distances. Further, participants in the U.S. (β = -0.03, 95% CI [-0.05, -0.01], p<0.001), Indonesia (β = -0.06, 95% CI [-0.08, -0.05], p<0.001) and Bangladesh (β = -0.02, 95% CI [-0.02, -0.01], p<0.001) were less closely genetically related to individuals at greater social distances.

Another concern in Bangladesh is that of a "floor" effect on amount foregone. Specifically, if Bangladesh participants had been offered a chance to sacrifice even less than 1/2kg rice to give their partner 5kg rice, we may have detected some effect of social distance. However, 1/2kg rice is already a very low level of sacrifice (10% of what the

partner would have received), and there was ample room for examining variation above that low level of potential sacrifice. Moreover, even Indonesian participants who had much higher levels of sacrifice did not exhibit social discounting, independent of other effects (e.g., relative need).

It is possible that the effect of social discounting only exists when participants interpret decisions as independent, whereas Bangladeshi and Indonesian participants tended to interpret the task in a cumulative context. This would represent a previously-unknown boundary condition for social discounting. To assess this possibility, we tested for an interaction effect of participant consistency and social distance on expected sharing, using a multilevel model with random effects for each individual. Because all Indonesian participants were inconsistent, this analysis is limited to the U.S. and Bangladesh. Among participants with non-zero generosity, there is no evidence that inconsistent participants exhibited less social discounting than consistent participants in the U.S. (β = -0.01, 95% CI [-0.09, 0.06], p = 0.687) or Bangladesh (β = -0.01, 95% CI [-0.08, 0.06], p = 0.760) (See Appendix for BIC-approximated Bayes Factors). This suggests that a cumulative interpretation of the task is not sufficient to explain a lack of social discounting.

It remains possible that the current design was insufficiently powered to detect a true relationship between social distance and generosity. To assess this possibility, we used the SIMR (Green & MacLeod, 2016) package in R (R Core Team, 2017) to estimate the effect size that our study had 95% power (α = 0.05) to detect. We find that the current study had approximately 95% power to detect an effect of [β = -0.018] in

Bangladesh and [β = -.040] in Indonesia. This indicates that the current design was sufficiently powered to detect all but the smallest effects. Detecting substantially smaller effects would require samples that are orders of magnitude larger than those of the current study. To illustrate, we calculate the number of participants needed to have 95% power (α = 0.05) to detect a true relationship between social distance and generosity, using the estimated effect sizes from Table 1. For Bangladesh, detecting an effect of [β = 0.002] would require approximately 3200 participants. For Indonesia, detecting an effect of [β = -0.006] would require approximately 4200 participants.

Discussion

We adapted a common social-discounting protocol, using real stakes, for implementation in 3 diverse populations: rural Bangladesh, rural Indonesia, and U.S. undergraduates. U.S. undergraduates displayed typical patterns of social discounting, replicating findings from numerous previous studies: participants incurred substantially greater costs to benefit socially-closer individuals. However, we also found a fundamental difference between U.S. undergraduates and the two rural populations in the relationship between social distance and generosity. In stark contrast, Bangladeshi and Indonesian participants did not exhibit social discounting. Further, participants in all sites were more generous to partners categorized as having greater relative need. These findings were consistent with respondents' post-decision rationales for their choices and could not be explained by several potential methodological concerns.

Our protocol differed in several ways from typical protocols (see https://osf.io/cfkdr/; (Hruschka et al., in press)). These modifications were necessary to

implement a study in rural settings with lower-literacy rates and participants unfamiliar with typical economic games, and we do not yet know how they affected our results. The fact that our findings among U.S. participants were consistent with past U.S. findings provides evidence that our study retained key aspects of typical protocols. And the fact that we documented the same pattern in Bangladesh and Indonesia (i.e. need predicts generosity but social distance does not) despite different protocols (See Methods) provides convergent evidence that these findings are not artefacts of one specific method (Munafò & Smith, 2018). Nonetheless, all operationalizations are imperfect and even seemingly arbitrary differences between protocols can generate dramatically different results (Landy & others, n.d.). We know embarrassingly little about what construct is captured by the typical measure of generosity used in this study (dichotomous choices between amounts of currency for self-versus other) and how well it correlates with different alternative operationalizations. This problem is not unique to social discounting (see (Frey et al., 2017) for a similar issue in studies of risk-preference) and is an important direction for future research.

Despite working with informants to ensure appropriate translations, it is possible that key concepts (e.g. social closeness) were understood differently by participants in Indonesia and Bangladesh than U.S. undergraduates. However, at least in Bangladesh, extensive interviewing suggests that the local term, *ghonishto*, is the appropriate modifier to describe relationships as intimate, close, or familiar. In interviews about their *ghonishto* friends and relationships, people mentioned that one helps them, can rely on them for help, and can talk with them about sensitive matters. In Indonesia, translators

and local research assistants identified *dekat* as the appropriate term to describe close and intimate relationships. Further, in all three sites, social distance was negatively correlated with another measure of closeness (i.e. IOS) and genetic relatedness (see results). This provides evidence that the idioms used in Bangladesh and Indonesia were roughly comparable to U.S. meanings.

Our study absolutely does not support general claims such as "humans are not more generous to socially-close partners" or "people in Bangladesh are not generous". Rather, it provides one piece of evidence against the hypothesis that social discounting is a cross-culturally robust phenomenon. Our study also provides evidence against the hypothesis that generosity is hyperbolically related to social distance (Jones & Rachlin, 2006): if social distance is unrelated to generosity, then this precludes a hyperbolic relationship with generosity. We can only speculate as to why our results diverge from prior findings. One plausible hypothesis is differences in cultural norms across sites. For example, informal interviews in Bangladesh revealed that giving without recipient need is a frowned-upon signal of superiority. Norms about whether people should behave according to personal preferences versus formal social obligations also vary across cultures, and it is possible that social discounting only exists when people treat others based primarily on individual feelings (Miller & Bersoff, 1998). It is also possible that populations in resource-scarce environments have norms that encourage a focus on relative need over other factors. Future research can assess this possibility by implementing social-discounting protocols among populations that vary in resource scarcity.

Given that norms strongly shape human behavior across societies, determining their precise effect on social discounting is a promising direction for future research. One approach is to manipulate experimental framing (e.g. instructing participants to make decisions based on "your duties or obligations towards this person" vs "your personal feelings towards this person"), as slight changes in framing can dramatically affect behavior and cognition in cross-cultural settings. However, when norms are internalized, slight framing-changes may be insufficient to change behavior. In such cases, comparative studies in diverse cultural and ecological settings may be our main window into the scale of human diversity (Henrich et al., 2005).

Towards a science of reliable and general phenomena

Our cross-cultural investigation of social discounting serves as one case study to illustrate the importance of checks on generalizability across diverse populations as a complement to narrowly-focused replication efforts. For any phenomenon, we should strive to conduct those studies that are most valuable to the scientific community. Just as some studies will not be worth replicating (Brandt et al., 2014; Coles, Tiokhin, Scheel, Isager, & Lakens, 2018), some will not warrant checks on generalizability (e.g. if an effect has weak empirical support, has a weak theoretical foundation, or does not even hold up to different operationalizations of theoretical constructs). But checks on generalizability are often essential. As this paper demonstrates, studying phenomena in previously-unstudied populations can be useful. However, distinct circumstances warrant distinct approaches to assessing generalizability (Apicella & Barrett, 2016). If we hypothesize that a reliably-replicated phenomenon is universal, we should strive to test it

among a diverse set of human populations. If we hypothesize that a phenomenon is unique to populations that exhibit some trait (e.g. that only populations with languages that use number words can count exact numbers of large magnitudes), then we can test this hypothesis by using a few target societies as critical tests (e.g. comparing performance on counting tasks in societies with and without number words) (Apicella & Barrett, 2016). And if we hypothesize that a phenomenon varies as a function of some parameter (e.g. generosity as a function of economic deprivation), but are agnostic as to the source of this variation (e.g. between cultures; between individuals within the same culture), then both within and between-culture studies can be useful (e.g. economic games among individuals from differentially-affluent neighborhoods within the same city; economics games among individuals from countries with different levels of affluence) (Nettle, 2017).

Without seriously considering how to improve the generalizability of our science, we put ourselves at risk. Foremost, we risk generating narrowly-replicable effects and theories of human behavior that tell us little about humanity as a whole (Crandall & Sherman, 2016; Henrich et al., 2010; Rozin, 2009). But we also face another risk: failing to study generalizability in a way that maximizes the scientific value of our research. Cross-cultural studies are costly, and it is often difficult to recruit large numbers of participants (especially for lone field researchers working in small-scale societies). Two unfortunate consequences are that many behavioral and psychological studies of non-WEIRD populations rely on small numbers of participants and are never directly replicated. Documenting cross-cultural variation may be a necessary first step towards

developing generalizable theories of human nature, but it will not be sufficient unless we invest in complementary efforts to ensure that such variation is reliable.

We see several potential ways to address these concerns. One is to increasingly invest in long-term field sites in non-WEIRD contexts (e.g. (Gurven et al., 2017)) where it is more feasible to acquire large sample sizes and conduct follow-up research. Another is to invest in large-scale collaborative projects that investigate the same phenomenon in diverse contexts (Barrett et al., 2016; Bryant et al., 2016; Henrich et al., 2006). One laudable recent initiative, The Psychological Science Accelerator (PSA), has taken the latter approach by developing a distributed network of laboratories across more than 50 countries (Moshontz et al., 2018). The PSA has tremendous potential: cross-cultural research projects will be able to acquire previously inconceivable sample sizes and lab-based experiments could plausibly allow better assessment of how contextual variables influence effect heterogeneity (Moshontz et al., 2018). Nonetheless, the PSA also faces major challenges. For example, if studies rely primarily on college undergraduates in different cultures, they will inevitably recruit participants who are wealthier, more educated, live in settings that more urban and industrialized, and are otherwise unrepresentative of much of humanity. This could result in a misleading picture of human diversity: consistent findings across different labs may be interpreted as establishing universality, whereas an effect actually depends on a parameter that does not sufficiently vary between labs (e.g. exposure to modern society, (Apicella, Azevedo, Christakis, & Fowler, 2014)). The extent to which this will be a major issue remains to be determined.

In our continual search for ways to improve our science, we should strongly consider the benefits of increasing investment in tests of generalizability across diverse populations. Although this point is not new (Crandall & Sherman, 2016; Henrich et al., 2010; Rozin, 2009), we believe it deserves far more attention in current discussions on scientific reform. Our plea for stronger checks on generalizability does not imply that replication is not important. On the contrary, an exclusive focus on exploring generalizability without direct replication risks generating a range of interesting effects that are difficult to explain and have questionable reliability (Zwaan et al., 2018). Improving our science requires investment in both determining the reliability of effects and assessing their generalizability across diverse contexts, cultures, and populations. Although norms and incentives are shifting in favor of the former, the latter remains woefully undervalued. This needs to change. Only by doing so will we develop models of human nature that are both reliable and broadly relevant.

Constraints on Generality (COG)

Constraints on Generality (COG) Statement (Simons et al., 2017a). The current study found no social discounting among participants from rural Bangladesh and Indonesia. We were surprised by this finding, and can only speculate as to the conditions in which it will replicate. *Participants.* Social distance and generosity: rural, resource-scarce populations in Asia. We have no reason to believe that the effect of recipient need on participant generosity depends on other characteristics of participants. *Materials.* We have no reason to believe that the results depend on characteristics of the specific materials used in our social-discounting protocol, although the lottery procedure may

increase the likelihood of "inconsistent" responding. *Procedures.* Social-closeness should be translated using the same terminology as the current study (i.e. *ghonishto* in Bangladesh; *dekat* in Indonesia). Participants should pass a comprehension check before starting the social-discounting protocol, and be ensured that their decisions are anonymous. *Historical/Temporal Specificity.* The effect of need on generosity may be driven by cultural norms that promote helping individuals with greater financial need. If so, this effect should not occur when such norms do not exist, or when norms promote exploiting individuals with greater financial need. The effect of social distance on generosity may be affected by cultural norms that sanction giving without need. If so, this effect should not occur when such norms exist. We have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

Ethics

Permission to perform this study was granted by the Arizona State University Institutional Review Board: STUDY00004602; STUDY00001752, 1201007249. All participants provided informed consent.

Data accessibility

All data, materials, and code are openly available at the Open Science Framework (https://osf.io/cfkdr/). The review of prior social-discounting research was pre-registered (https://osf.io/mw37t/). The direct replication of our lab's prior social-discounting study was pre-registered (https://osf.io/fn9am/). The social discounting study and coding of participant responses were not pre-registered.

CHAPTER 2

REGISTERED REPORT: AN EXPERIMENTAL TEST OF THE EFFECTS OF COMPETITION FOR PRIORITY ON INFORMATION SAMPLING

Abstract

**This study has received in-principle acceptance as a Registered Report in Royal Society Open Science.*

Scientific progress depends on the reliability of scientific findings. Yet, recent failures to replicate findings in several fields demonstrate that published literature is often unreliable. Although many factors plausibly affect the reliability of scientific findings, incentive structures are thought to be fundamental. One longstanding incentive, rewarding priority of publication, may have the negative effect of incentivizing rushed, low-quality research. Here we develop a laboratory experiment to investigate how competition affects information sampling in a game that parallels scientific investigation. Individuals must gather data in order to guess true states of the world, are incentivized to make as many correct guesses as possible, and face a tradeoff between guessing quickly and increasing accuracy by acquiring more information. To test whether competition affects accuracy, we compare a condition in which individuals are rewarded for each correct guess to a condition where individuals face the possibility of being "scooped" by a competitor that makes the correct guess more quickly. In a second set of experimental treatments, we make information harder to acquire by making information-acquisition contingent on solving arithmetic problems. This allows us to test whether competition necessarily causes individuals to trade accuracy for speed or whether individuals can

avoid this tradeoff by increasing the rate at which they acquire information. In doing so, this study takes one step towards understanding how competitive incentives affect information-sampling strategies.

Introduction

A central aim of science is generating reliable results to produce increasingly accurate theories about the world. However, reliability can't be taken for granted. Models of the scientific process suggest that, given current scientific practices, many research findings will be false (Ioannidis, 2005; Richard McElreath & Smaldino, 2015; Nissen, Magidson, Gross, & Bergstrom, 2016). For example, publication bias in favor of positive results and combined with low statistical power is predicted inflate the ratio published false positive to positive results (Button et al., 2013). Empirical findings from large-scale replication efforts in diverse fields are consistent with these prediction: many results fail to replicate (Begley & Ioannidis, 2015; Camerer et al., 2016; Collaboration & others, 2015; Cova et al., 2018; Nosek & Errington, 2017). Science may be our most powerful tool for generating knowledge, but there is clearly room for improvement.

Many factors plausibly affect the reliability of science (Munafò et al., 2017). At the heart of these are incentive structures: by determining the professional payoffs for various types of research, incentives shape scientists' research decisions (Nosek, Spies, & Motyl, 2012). Many current incentives are thought to harm the efficiency of science, including publication bias in favor of positive and novel findings, evaluating scientists based on number of publications, and lack of rewards for data sharing and transparent research (Munafò et al., 2017). In recent years, scholars have been especially concerned

about the harmful effects of competition on the scientific process (Alberts, Kirschner, Tilghman, & Varmus, 2014; Anderson, Ronning, De Vries, & Martinson, 2007; Benedictus, Miedema, & Ferguson, 2016; Fang & Casadevall, 2015; Geman & Geman, 2016; Nosek et al., 2012; Rawat & Meena, 2014; Sarewitz, 2016, 2016; Smaldino & McElreath, 2016). Competitive incentive structures are circumstances in which individuals can expend finite resources (e.g. time, money) to increase their probability of receiving a payoff, and one individual's success reduces the probability that others will succeed (Dechenaux, Kovenock, & Sheremeta, 2015). By this definition, much of science is competitive: scientists compete for publication in journals, limited professional positions, and funding opportunities (Anderson et al., 2007; Balietti, Goldstone, & Helbing, 2016; Fang & Casadevall, 2015; Nosek et al., 2012).

Competition over priority of discovery is arguably one of the most important forms of competition in science. Academic science has a longstanding norm of rewarding individuals for making discoveries and publishing novel findings. Over 50 years ago, sociologist of science Robert Merton noted how this norm might benefit science: rewarding priority can incentivize scientists to invest effort to quickly solve important problems and share their discoveries with the scientific community (Merton, 1957). Models of academic priority races substantiate Merton's intuition: under some conditions, rewarding priority of discovery can incentivize the disclosure of partial results (Banerjee, Goel, & Kollagunta Krishnaswamy, 2014; Bergstrom, Foster, & Song, 2016; T. Boyer, 2014; Heesen, 2017) and lead to efficient distributions of scientists across research problems (Strevens, 2003). Nonetheless, scholars have also had longstanding

concerns about the repercussions of this norm. Charles Darwin thought that rewarding priority by naming species after their first-describers incentivized biologists to produce "hasty and careless work" by "miserably describing a species in two or three lines " (21, p. 644). More recently, concerns over the consequences of rewarding priority have led the academic journals eLife and PLOS Biology to offer "scoop protection" (i.e. allowing researchers to publish findings identical to those already published in the same journal) in attempts to reduce the disproportionate payoffs to scientists who publish first (Marder, 2017; The PLOS Biology Staff Editors, 2018; Yong, 2018b). In the editorial justifying their new policy, The PLOS Biology Staff Editors write "…many of us know researchers who have rushed a study into publication before doing all the necessary controls because they were afraid of being scooped. Of course, healthy competition can be good for science, but the pressure to be first is often deleterious…" (The PLOS Biology Staff Editors, 2018).

Despite these reasonable concerns, there is little empirical evidence for the hypothesis that competitive pressures to publish cause individuals to produce lower-quality research. In focus-group discussions with mid and early-career researchers, scientists acknowledge that competition incentivizes them to conduct careless work (Anderson et al., 2007), but laboratory experiments investigating competition more broadly demonstrate that competition also promotes individual effort (Baer, Vadera, Leenders, & Oldham, 2013; Balietti et al., 2016; Dechenaux et al., 2015; Gneezy, Niederle, & Rustichini, 2003). As a consequence, it is unclear how competition in general, and competition for priority in particular, affects research quality. On the one

hand, competition might cause researchers to make dubious claims based on inadequate data. On the other, competition might encourage researchers to gather data more efficiently.

Given the difficulty of experimentally manipulating incentives in real-world scientific practice, we developed a simple game that mimics aspects of scientific investigation. In our experiment, participants must gather data in order to guess true states of the world and face a tradeoff between guessing quickly and increasing accuracy by acquiring more information. Although this game is a simplification of the scientific process, leaving out many factors that exist in real-world scientific research, it allows us to investigate two hypothesized effects of competition on information-sampling strategies in controlled conditions. By doing so, our experiment brings quantitative data to the debate about whether competition necessarily causes individuals to sacrifice research quality by trading accuracy for speed or whether individuals can avoid this tradeoff by modulating their effort.

Study Aims

We develop a simple experiment to test the effect of competition for priority on information-acquisition strategies. To do so, we modify the Cambridge Information Sampling Task (IST) (Clark, Robbins, Ersche, & Sahakian, 2006) to create a game in which individuals gather information in order to guess true states about the world. Players are incentivized to make as many correct guesses as possible and face a tradeoff between guessing quickly and increasing accuracy by gathering more information.

In order to investigate the effect of competition on how individuals acquire information, we compare a baseline treatment in which players are rewarded for each correct guess to a treatment where players face the possibility of being "scooped" by a competitor that makes the correct guess more quickly. In a second set of treatments, we make the rate at which individuals can acquire information contingent on individuals' effort. Doing so allows us to test whether competition to guess first leads to a faster rate of information acquisition.

This set of treatments investigates two potential effects of competition on information acquisition: a negative effect on reliability (individuals might make inferences from smaller amounts of evidence) and a positive effect on productivity (individuals might work to acquire evidence at a faster rate). Such a design allows us to test whether competition necessarily encourages individuals to trade accuracy for speed or whether individuals can avoid this tradeoff by adjusting their level of effort.

Below, we outline the experimental design and hypotheses, and present a simple analytical model based on the experimental parameters. We then outline all planned analyses and present results from a pilot study.

Game Design

This computer game (a modified version of the Cambridge Information Sampling Task (Clark et al., 2006); programmed in Object Pascal with Delphi 7) provides a simple instantiation of a process in which individuals decide how much information to gather when solving a problem and face tradeoffs between quickly producing an answer and increasing accuracy via larger samples. Players view a screen with 25 black tiles arranged

in a 5 x 5 grid (Figure 1a). Each tile has one of two underlying colors (yellow or blue) and participants can click any tile to reveal its underlying color. Players must guess the grid's majority (i.e. most common) color and are rewarded for accurate guesses. After guessing, players move on to the next grid. Players are informed whether their guess was correct or incorrect and can see their cumulative score. In the No-Competition treatments (see Treatments) participants play 20 minutes and so face a speed-accuracy tradeoff: guessing earlier (i.e. with few tiles clicked) allows them to quickly move on to subsequent grids but decreases their probability of guessing correctly.

The proportion of yellow and blue tiles and their order of appearance are deterministic but remain unknown to players. This controls for stochasticity in access to information by ensuring that each player receives the same information in the same order, regardless of the tiles that a player clicks. Each grid is characterized by a proportion of yellow and blue tiles (i.e. effect size). This proportion is chosen randomly from one of three possible ratios (8:17, 10:15, 12:13) and yellow and blue tiles have the same probability of being in the majority. To control for stochasticity in the grids solved by different participants, all participants receive the same grids in the same order. Because one of the two colors might be more salient for the human visual system (e.g. 4 blue: 2 yellow might be a stronger visual signal for blue than 4 yellow: 2 blue is a signal for yellow), the baseline color is randomly selected at the beginning of the experiment for each participant (e.g. some participants see Y-Y-B-Y-…-Y and have to guess Y, while others see B-B-Y-B-…-B and have to guess B). This aims to limit unforeseen bias.

Treatments

We use a 2 x 2 between-subjects design to investigate two treatments (No Competition; Competition) and two conditions (No Effort, Effort).

No-Effort Condition

No Competition: Participants play the game for 20 minutes and can acquire information by clicking 1 tile every 1 second. The 1-second delay between clicking tiles prevents players from increasing their clicking speed to acquire information faster than 1 tile per second. Players receive payoffs solely as a function of their own performance: they gain 1 point for each correct guess and lose 1 point for each incorrect guess. Players must wait for 5-seconds between making their guess and being presented with the next grid. This payoff structure incentivizes players to acquire at least some information before guessing, because guessing without clicking any tiles results in a 50% probability of answering correctly and hence an expected payoff of 0.

Competition: Players compete against the performance of one previous same-sex participant from the No-Competition treatment. These competitors will be sampled without replacement: each participant will compete against the performance of one unique previous participant. After submitting their guess, players move on to the subsequent problem, where they again compete against the same player's performance on that problem until they solve the same set of grids as their competitor. For grids where the competitor guesses correctly, players receive a payoff only if they are correct and guess faster than their competitor. For grids where the competitor guesses incorrectly, players receive a payoff solely as a function of their own performance. Players are informed of their payoff at the end of each round and are notified whenever they guess

after a competitor who has already guessed correctly. Players are not informed about the number of tiles revealed by their competitor.

We instantiate competition by having players compete against the performance of baseline participants for several reasons. A more realistic instantiation would be for two players in the competition treatment to directly compete against each other. In this design, for both players to remain synchronized, they would need to move to the next grid as soon as either player guessed. Such a design is problematic: it only generates data from 1 player per grid (i.e. the guesser) and that data is biased towards the player that reacts most to competition by guessing earliest. One potential solution is to synchronize players by allowing both players to make their guess on each grid. However, this introduces other problems: the player that guesses earlier is forced to wait for their competitor to guess. As a consequence, the two competitors would not experience the exact same experimental conditions, which may result in experimental biases. For example, players might delay their guesses to avoid waiting for their competitor to guess. Our design avoids these problems.

Effort Condition

Treatments in the Effort condition are identical to treatments in the No-Effort condition except that participants need to solve a simple arithmetic problem before being able to click on a tile (Figure 1b). This is a commonly used real-effort task in economics (Lezzi, Fleming, & Zizzo, 2015). In this treatment, players can increase their rate of clicking tiles by solving arithmetic problems more quickly (See Pilot Study for checks on floor effects).

a. **b.**

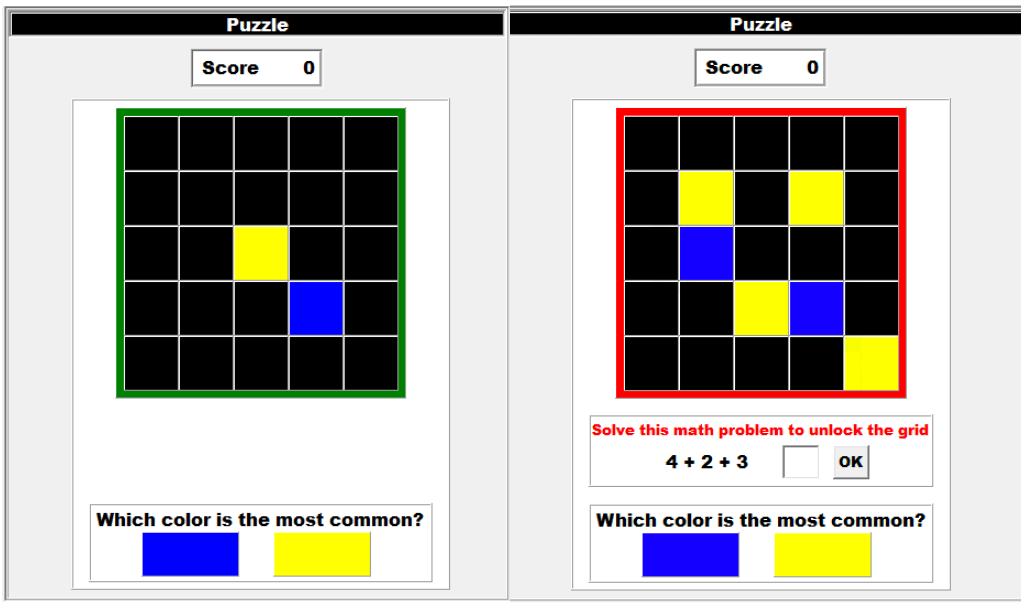


Figure 1| Game Principle. The experimental task consists of 25 black tiles arranged in a 5 x 5 grid. Players can click on tiles to reveal their underlying color (yellow or blue) and are rewarded for correct solutions (i.e. correctly guessing the majority color). Clicking more tiles provides more information about the grid but takes more time. The order in which colors are revealed is deterministic (i.e. independent of which tile is clicked) but remains unknown to players. (a) In the No-Effort condition, players can click a black tile every 1 second. (b) In the Effort condition, players need to solve an arithmetic problem to acquire information (i.e. clicking a tile). In the No-Competition treatment, players gain/lose 1 point for each correct/incorrect guess. In the Competition treatment, players only gain/lose points if they guess sooner than their competitor.

Sampling Plan

Arizona State University students (equal numbers of women and men, age 18 and over) will be randomly selected from a database managed by the Elinor Ostrom Multi-Method Lab at Arizona State University and recruited by email. We will obtain informed consent from all subjects before the experiment's onset (ethical approval has been obtained from the Arizona State University IRB, code: STUDY00007691). Participants

will receive $5 for participation and an additional amount ranging from $0 to $10 depending on their performance (see Score Calculation).

Procedure

The experiment will take place in a computer room at Arizona State University. Each session will consist of a maximum of 16 participants (exclusively male or female) who will all be assigned to either a Competition or No-Competition treatment. Participants in the Competition treatments will be randomly assigned to compete against the performance of one same-sex player from the equivalent No-Competition treatment. This requires that the first session be the No-Competition treatment. Subsequent sessions will alternate between Competition and No-Competition. Within each session, each participant will be randomly assigned to either the Effort or No-Effort condition. All manipulations (Competition vs No Competition; Effort vs No Effort) are between subjects.

Participants will enter the computer room in the order that they arrive to the experiment, will sit at physically-separated computers, and will be instructed that communication is not allowed. Participants will be blind to the fact that there are four experimental treatments. Before starting the experiment, participants will be requested to enter their age and sex. Participants will be shown the instructions on their screens. In the No-Competition treatments, the game will last 20 minutes. In the Competition treatments, the game will last as long as it takes players to complete the same set of grids that were completed by their competitor. At the end of the game, participants will receive a reward according to their performance (see Score Calculation).

Tutorial and Pre-Game Information

Before starting, players will complete a tutorial in which they will perform basic actions. This tutorial will guide players' actions so that players experience clicking tiles and choosing the majority color, to ensure that all players have mastered the interface before playing. Players in the No-Competition treatments will be informed that the goal of the game is to make as many correct guesses as possible within the experiments' duration. Players in the Competition treatments will be informed that they are playing against the performance of another participant, and that the goal of the game is to be first to make as many correct guesses as possible. Players will be informed that their score and monetary reward will depend on their total number of correct guesses and correct faster-guesses, respectively. In the Effort condition, players will experience solving a math problem. Players will not be informed about the total number of yellow and blue tiles per round.

Score Calculation

In the No-Competition treatments, players will receive one point for each correct guess and will lose one point for each incorrect one. In the Competition treatments, the payoff structure is identical when participants guess a) sooner than their competitor, or b) after a competitor who guessed incorrectly. Participants do not gain or lose any points when guessing after a competitor who guessed correctly. This payoff structure corresponds to the assumption that being scooped prevents researchers from both receiving a benefit for being correct and from paying a cost for being wrong. In the unlikely event that a player guesses at the exact same time as their competitor, they will

receive one point. A player's final score will be the sum of these points. The function that translates scores into payoffs will remain unknown to players:

$$\text{Payoff}_{\text{no-competition}} = \$0.15 \times \text{CorrectGuesses} - \$0.15 \times \text{IncorrectGuesses}$$

$$\text{Payoff}_{\text{competition}} = \$0.15 \times \text{CorrectFasterGuesses} - \$0.15 \times \text{IncorrectFasterGuesses}$$

Data Collection Stopping Rules

We specify a region of practical equivalence (ROPE) for all relevant parameters (34, see "Analyses and Predictions" below). We will stop data collection after the 95% highest probability density interval (HPDI) for all parameters falls entirely inside or outside pre-specified ROPEs for each hypothesis. We will check whether the HPDIs fall inside or outside each ROPE after every 4 sessions of data collection (1 session corresponding to each of the 4 treatments; maximum 16 participants per session) and will collect data until we obtain a maximum of 260 participants (an upper limit set by funding availability). Data from participants excluded based on pre-specified criteria will not count towards this 260-participant limit.

Completion Timeline

If stage 1 review is successful, we anticipate completing the experiment and submitting the manuscript for stage 2 review within 5 months of receiving stage 1 approval.

Model

We developed a simple mathematical model to better understand the payoff structure of our experiment and to gain insight into how competition should affect players' behavior in the No-Effort conditions. The goal of this model is to understand

whether H1a (see Hypotheses below) is logically coherent (i.e. that our instantiation of competition for priority actually incentivizes participants to guess with smaller amounts of evidence).

In the experiment's Competition treatments, players who guess the underlying color only gain or lose 1 point if they a) take less time to guess than their opponent or b) they take longer to guess than their opponent, but their opponent has guessed incorrectly. When a player guesses before or at the same time as an opponent, the player's expected payoff (*EP*) is:

$$EP = p_p * 1 + \left(1 - \ p_p\right) * (-1)$$

$$EP = 2p_p - 1$$

where p_p is the probability that the player guesses correctly. When a player guesses after their opponent, the player's *EP* is:

$$EP = p_o * 0 + \ (1 - \ p_o) * \ p_p * 1 + \ \ (1 - \ p_o) * \ \left(1 - \ p_p\right) * (-1)$$

$$EP = 2p_p + \ p_o - \ 2p_p p_o \ - \ 1$$

where p_o is the probability that the player's opponent (i.e. the player in the no-competition treatment) guesses correctly. This assumes that guessing at the same time or before an opponent are equivalent, and that only guessing after an opponent results in some probability of being scooped. *EP* can take on values between 0 and 1 because players have a minimum 0.5 probability of correctly guessing the underlying color and payoffs to correct and incorrect guesses are symmetrical. p_p and p_o are a function of two parameters: the ratio of colored tiles (i.e. effect size 8:17, 10:15, or 12:13) and the number of tiles that a player/opponent reveals (we assume that players' time-to-guess is

entirely determined by the number of tiles they reveal). To calculate p_p and $p_{o,}$, we simulated the average amount of information available to a player, conditional on the effect size and the player revealing a given number of tiles (Fig. S1, code available at https://osf.io/udm8g/). For each effect size, we randomly generated 25-tile sequences of yellow and blue tiles (500 simulations). We then computed the proportion of all possible 25-tile sequences that give the same outcome as the current majority color after *n* tiles have been revealed. For example, consider a player who reveals a 7-tile sequence of Y-Y-Y-Y-B-Y-Y. Given this initial sequence of tiles, there are 230964 25-tile sequences, out of all possible sequences, that result in a majority yellow color and 31180 sequences that result in a majority blue color. The amount of information available to this player is 230964 / (31180 + 230964) = 0.881. This means that if the player guesses that yellow is the majority color from that specific 7-tile sequence, she will be correct 88.1% of the time.

Figure 2 plots *EP* as a function of the number of tiles that a player and their opponent reveal, for different effect sizes. If an opponent reveals many tiles, a player receives the highest *EP* by revealing the exact same number or slightly fewer numbers of tiles. This occurs because players have a high probability of guessing the majority color when they reveal a large number of tiles and guessing before an opponent guarantees that the player does not get scooped. If a player guesses after an opponent who has revealed many tiles, a player's *EP* is low: the opponent will usually correctly guess the majority color, causing the player to obtain 0 points. If an opponent reveals very few tiles, a player receives the highest *EP* by revealing a large number of tiles. This occurs because a

player who guesses before this opponent has a high probability of guessing incorrectly, whereas a player who guesses after this opponent can maximize their probability of correctly guessing the majority color by revealing as many tiles as possible. As effect sizes decrease, a player receives higher *EP* by revealing more tiles. This occurs because, for most possible numbers of tiles revealed below 25, a player has a lower probability of guessing correctly when effect sizes are small.

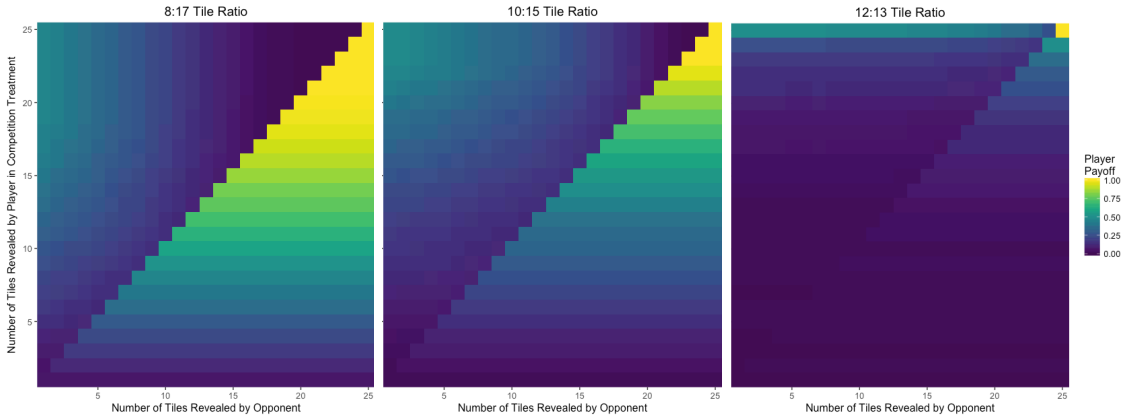


Figure 2| Player's expected payoff as a function of the number of tiles revealed by the player and their competitor. Plotted (left to right) for three ratios of colored tiles (i.e. effect sizes): 8:17, 10:15, and 12:13. X and Y-axes indicate the number of tiles revealed by a player's opponent and the player, respectively. Players' expected payoff is highest when a competitor reveals a large number of tiles and players reveal the exact same or fewer tiles. The largest drop in payoff occurs when players reveal slightly more tiles than a competitor. When a competitor guesses after revealing few tiles, players maximize their expected payoff by revealing many tiles before guessing.

To assess the logical coherence of H1a, we calculated players' *EP* as a function of the effect size and number of tiles revealed, assuming that players compete against payoff-maximizing competitors (i.e. against individuals who reveal the number of tiles that maximizes their expected-payoff in the No-Competition, No-Effort treatment). The results corroborate the intuition suggested by a visual inspection of Figure 2: when

competing against payoff-maximizing competitors (who reveal large numbers of tiles), players maximize *EP* by revealing the same number or fewer tiles than their competitor (See Appendix). Sensitivity checks indicate that this result is robust to incorporating stochasticity in the number of tiles revealed by competitors.

This model provides support for the logical coherence of H1a: the experiment's payoff structure incentivizes players to guess at the same time as or before their competitors. However, the former outcome is unlikely: in the experiment (unlike the model), priority is determined by amount of time spent before guessing, not by number of tiles revealed. As such, players in the Competition, No-Effort treatment usually maximize their expected payoff by guessing before competitors.

Hypotheses

Condition 1: No Effort

(H1a) Competition for priority will cause participants to guess with smaller amounts of evidence.

In the first set of experimental treatments, individuals cannot acquire information faster than a predetermined rate (See Figure 1a). As a consequence, when individuals are in competition to guess correctly before a competitor, they can only do so by making their guess with less information (i.e. removing less tiles before guessing).

(H1b) If H1a is confirmed, then competition for priority will also cause participants in Condition 1 to have reduced accuracy.

If participants remove less tiles before guessing, then they will have a lower probability of making a correct guess on each grid, because the probability of making a correct guess is a monotonically increasing function of the number of tiles revealed.

Condition 2: Effort

(H2) Competition for priority will increase participant effort, thereby causing participants to reveal information faster.

In the second set of experimental treatments, the rate at which information can be acquired depends on individuals' effort: individuals need to solve an arithmetic problem for each piece of information (see Figure 1B). Individuals can thus affect their rate of information acquisition by adjusting their effort (i.e. solving arithmetic problems faster or slower). Empirical research in experimental economics has found that competition increases participants' effort in similar tasks (Dechenaux et al., 2015; Lezzi et al., 2015). As such, we expect competition for priority to cause participants to solve arithmetic problems at a faster rate (H2).

Interaction between Condition 1 and Condition 2

(H3a) The effect of competition for priority on reducing the amount of evidence that participants gather will be larger in Condition 1 than in Condition 2.

(H3b) If H3a is confirmed, then competition will cause a bigger reduction in participant accuracy in Condition 1 than in Condition 2.

In the Effort condition, individuals can potentially guess before their competitor in two ways: reveal less information (i.e. fewer tiles) before guessing or increase effort

and solve arithmetic problems more quickly. These are not mutually exclusive. However, if individuals do increase their speed of solving arithmetic problems, then we expect competition to have a smaller effect on their accuracy. We make no prediction about whether competition will have absolutely no effect on participant accuracy in the Effort condition or whether competition will simply have a smaller, negative effect on accuracy, compared to the No-Effort condition.

Analyses and Predictions

We will fit statistical models within a Bayesian framework with weakly informative priors, using map2stan in Richard McElreath's *Rethinking* package in R (R. McElreath, 2012; R Core Team, 2017). Table 1 (see "Priors" below) lists prior probability distributions for parameters in all statistical models.

Power Analysis

To check whether 260 participants provides sufficient statistical power to evaluate our hypotheses, we used our pilot data (see "Pilot Study" below) to conduct a power analysis. For confirmatory analyses, we set a ROPE for each relevant parameter by determining the minimum effect size that could be detected 95% of the time, given our maximum sample size of 260 participants. This minimum effect size then determined the upper and lower bounds of the ROPE for each analysis (Lakens, 2017; Lakens, Scheel, & Isager, 2017). We considered an alternative approach: setting all ROPE boundaries based on theoretical considerations (J. K. Kruschke, 2018). However, because our hypotheses only make directional predictions, they provide no guidance as to an effect's size or minimum effect sizes of interest.

We collected data in the pilot study to be able to conduct power analyses for H1 (a, b) and H3 (a, b). For H2 (Model 3), we did not record participants' time between clicking one tile and being allowed to click subsequent tile (i.e. time to accurately solve one arithmetic problem) in the pilot study. As such, we instead conducted a power analysis for Model 3 using a slightly different outcome variable: time to produce any answer (accurate or inaccurate) for one arithmetic problem. We then set a ROPE based on 99% statistical power, to compensate for the uncertainty introduced by basing the power analysis on a statistical model with a different outcome variable.

For each power analysis, we followed the following steps (code available at https://osf.io/udm8g/):

1. Analyze the pilot data with Bayesian statistical models using map2stan in Richard McElreath's *Rethinking* package in R (see "Analyses and Predictions*" below) (R. McElreath, 2012; R Core Team, 2017).
    a. *Simulated data for Model 1 were assumed to be Gamma distributed (see code).
2. Extract samples for all parameters from the posterior probability distribution for a given statistical model.
3. Simulate 260 participants with *x* observations per participant, where *x* is randomly sampled (with replacement) from the number of observations per participant in the pilot.

4. Generate simulated data for each participant by taking random samples from the posterior probability distribution for each parameter and inserting those samples into the formula implied by the statistical model structure for an analysis.
5. Analyze the simulated data with a Frequentist implementation of the statistical model in the proposed analysis.
6. Record the confidence interval for each parameter of interest.
7. Repeat steps 2 – 6, 500 times.
8. Generate a ROPE for each parameter of interest by determining the maximum effect size that fell outside of the 95% of confidence interval in the 500 simulations.

Testing Hypotheses

For quality checks, we generated a ROPE for each parameter based on our subjective assessment of what effect size would convincingly indicate that a manipulation was successful. For confirmatory predictions, we generated a ROPE for each parameter by conducting power analyses to determine the minimum effect size that can be detected 95% of the time, given our maximum sample size.

If the 95% highest probability density interval (HPDI) for a parameter falls outside of the ROPE, we will consider this as evidence against the null hypothesis of no effect. If the 95% HPDI for a parameter falls outside of the ROPE and is in the direction predicted by a hypothesis, we will consider this as evidence for the hypothesis. If the 95% HPDI for a parameter falls outside of the ROPE and is in the opposite direction to that predicted by a hypothesis, we will consider this as evidence against the hypothesis. If

the 95% HPDI falls within the ROPE, we will consider this as evidence for the null hypothesis of no effect. If the 95% HPDI does not fall entirely within or outside the ROPE, we will consider that the study does not provide conclusive evidence for either the prediction or the null hypothesis.

Exclusions and Outliers

We will exclude all data from participants who did not complete the study (i.e. who did not answer the final "*Competition Attention Check*" question; see Quality Checks below). Within individual participants, we will exclude observations for which there is missing data for at least one measured variable. Both participants' time to make a guess and time to solve arithmetic problems follow heavily right-skewed distributions (see "Exclusions and Outliers" in Pilot Study). For participants' time to make a guess, we will exclude times that are more than 5 standard deviations larger than the mean time until making a guess. For participants' time to solve arithmetic problems, we will exclude arithmetic-problem solving times that are more than 5 standard deviations larger than the mean arithmetic-problem solving time. These criteria allow for exclusion of the most extreme data points while preventing exclusion of too many observations. These same exclusion criteria are also used in the analysis of the pilot data (see Appendix).

Quality Checks

Effort Manipulation: If the effort manipulation is successful, then participants in the Effort conditions should take longer per click than participants in the No-Effort conditions. To assess the effect of effort on the average time to click a tile, we will use a linear regression, with random effects for player, of the following form:

$$Yi \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \alpha_{\text{PLAYER}[i]} + \beta_E * E_i + \beta_C * C_i + \beta_{CE} * C_i E_i$$

$Y_{i:}$ Time (seconds) to click one tile and reveal its underlying color. $\alpha_{:}$ Intercept. $\alpha_{\text{PLAYER}[i]:}$ Random intercept for each player. E: Effort Condition (1 / 0). $C_{:}$ Competition Treatment (1 / 0). $\beta_{CE} * C_i \text{E}_{i:}$ Interaction between treatment and effort.

Competition Attention Check: At the end of the experiment, participants will be asked "During the experiment, were you competing with another player to be first to guess the correct answer?". If the participants in the competition treatment are aware that they competed against another individual, then a higher proportion of participants in the Competition treatments should answer "yes" to this question than in the No-Competition treatments. To assess the effect of competition on answering "yes" to this question, we will use a logistic regression of the following form:

$$Y_i \sim \text{Binomial}(1, p_i)$$

$$\text{Logit}(p_i) = \alpha + \beta_C * C_i$$

$Y_{i:}$ Answered "yes". $\alpha_{:}$ Intercept. $C_{:}$ Competition Treatment (1 / 0).

Confirmatory Analysis Plans

(H1a). Competition for priority will cause participants in the No Effort condition to guess with smaller amounts of evidence.

(H3a). Competition for priority will cause a bigger reduction in the amount of evidence that participants gather in the No Effort condition than in the Effort condition.

We will use one dependent measure to test whether competition causes players to guess with smaller amounts of evidence: the number of tiles that players reveal when making a guess.

In the Competition treatments, players should reveal fewer tiles when making their guess than in the No-Competition treatments. There should also be a positive interaction between the Competition treatments and the Effort condition (see Hypotheses above): the effect of competition on number of tiles revealed should be smaller in the Effort condition than in the No-Effort condition (due to the expected positive effect of competition on individual effort).

Model 1: To compare the number of tiles that players reveal before guessing the majority color in the Competition treatments versus the No-Competition treatments, we will use a multiple linear regression, with random effects for player, of the following form:

$$Yi \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \alpha_{\text{PLAYER}[i]} + \beta_C * C_i + \beta_E * E_i + \beta_{CE} * C_i E_i + \beta_{Ns} * Ns_i$$

Y_i: Number of tiles clicked before guessing. α: Intercept. $\alpha_{\text{PLAYER}[i]}$: Deviation from intercept for each player. C: Competition Treatment (1 / 0). E: Effort Condition (1 / 0). $\beta_{CE} * C_i E_i$: Interaction between treatment and effort. β_{Ns}: Standardized number of tiles for the majority color (i.e. effect size).

(H1b). If H1a is confirmed, then competition for priority will also cause participants in the No Effort condition to have reduced accuracy.

(H3b). If H3a is confirmed, then competition for priority will cause a bigger reduction in participant accuracy in the No Effort condition than in the Effort condition.

We will use one dependent measure to test the effect of competition for priority on accuracy: probability of correctly guessing the majority color. In the Competition treatments, players should have a lower probability of making correct guesses than in the No-Competition treatments. There should also be positive interaction between the Competition treatments and the Effort condition: the effect of competition on accuracy should be smaller in the Effort condition than in the No-Effort condition.

Model 2: To assess the probability of a correct guess, we will use a logistic regression, with random effects for player, of the following form:

$$S_i \sim \text{Binomial}(1, p_i)$$

$$\text{Logit}(p_i) = \alpha + \alpha_{\text{PLAYER}[i]} + \beta_C * C_i + \beta_E * E_i + \beta_{CE} * C_i E_i + \beta_{Ns} * Ns_i$$

S_i: Successful guess. α: Intercept. $\alpha_{\text{PLAYER}[i]}$: Deviation from intercept for each player. C: Competition Treatment (1 / 0). E: Effort Condition (1 / 0). $\beta_{CE} * C_i E_i$: Interaction between treatment and effort. β_{Ns}: Standardized number of tiles for the majority color (i.e. effect size).

(H2). Competition for priority will increase participant effort, thereby causing participants to reveal information faster.

We will use one dependent measure to test the effect of competition for priority on effort: amount of time (seconds) between when players reveal a piece of information (i.e. click a tile) and when players are able to click the next tile (i.e. the time to accurately

solve an arithmetic problem). This analysis will be limited to players in the Effort condition.

Players in the Competition X Effort treatment should solve arithmetic problems faster than players in the No-Competition X Effort treatment. The time that it takes players to solve arithmetic problems is the time between clicking one tile and being allowed to click the subsequent tile. As a result, players in the Competition X Effort treatment should have a smaller time between clicking one tile and being allowed to click the subsequent tile compared to players in the No-Competition X Effort treatment.

Model 3: To test the effect of competition on the time between clicking one tile and being allowed to click the subsequent tile (i.e. time to accurately solve one arithmetic problem), we will use a multiple linear regression, with random effects for player, of the following form:

$$Yi \sim \text{Normal}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \alpha_{\text{PLAYER}[i]} + \beta_C * C_i + \beta_{Ns} * Ns_i$$

Y_i: Time between clicking one tile and being allowed to click the subsequent tile (i.e. time to solve an arithmetic problem; seconds). α: Intercept. $\alpha_{\text{PLAYER}[i]}$: Deviation from intercept for each player. C: Competition Treatment (1 / 0). β_{Ns}: Standardized number of tiles for the majority color (i.e. effect size).

Priors

Parameter	Effort Manipulation Check	Competition Attention Check	Model 1 (Tiles)	Model 2 (Correct Guess)	Model 3 (Arithmetic Time)
σ	Gamma (2, 0.5)	NA	Gamma (2, 0.5)	NA	Gamma (2, 0.5)
α	Gamma (1.5, 0.05)	Normal (0, 10)	Uniform (0, 25)	Normal (0, 10)	Gamma (1, 0.05)
α_{PLAYER}	Normal (0, σ_{PLAYER})	NA	Normal (0, σ_{PLAYER})	Normal (0, σ_{PLAYER})	Normal (0, σ_{PLAYER})
σ_{PLAYER}	Gamma (1.5, 0.05)	NA	Gamma (1.5, 0.05)	Gamma (1.5, 0.05)	Gamma (1, 0.05)
B_C	Normal (0, 10)	Normal (0, 10)	Normal (0, 10)	Normal (0, 10)	Normal (0, 10)
β_E	Normal (0, 10)	NA	Normal (0, 10)	Normal (0, 10)	NA
β_{CE}	Normal (0, 10)	NA	Normal (0, 10)	Normal (0, 10)	NA
β_{Ns}	NA	NA	Normal (0, 10)	Normal (0, 10)	Normal (0, 10)

Table 1 | Priors for statistical models. Prior probability distributions for all statistical models, including quality checks ("Effort Manipulation Check", "Competition Attention Check") and confirmatory analysis plans (Models 1 – 3). Gamma distributions are defined by parameters for shape and rate. Normal distributions are defined by parameters for mean and standard deviation.

ROPEs

Parameter	Effort Manipulation Check	Competition Attention Check	Model 1 (Tiles)	Model 2 (Correct Guess)	Model 3 (Arithmetic Time)
B_C	NA	(-0.8, 0.8)	(-1.22, 1.22)	(-0.19, 0.19)	(-0.33, 0.33)**
β_E	(-0.5, 0.5)	NA	NA	NA	NA
β_{CE}	NA	NA	(-0.10, 0.10)*	(-0.09, 0.09)	NA

Table 2 | Region of practical equivalence (ROPE). ROPEs for quality checks are based on subjective assessment of what effect size would convincingly indicate a successful manipulation. ROPEs for confirmatory analyses (Models 1 - 3) are based on 95% statistical power, unless indicated otherwise. Model 1 tests the effect of the Competition treatment and Effort condition on number of tiles clicked before guessing, using a multiple linear regression with random effects for each player. Model 2 tests the effect of the Competition treatment and Effort condition on the probability of a correct guess, using a logistic regression with random effects for each player. Model 3 tests the effect of the Competition treatment on the time to accurately solve one arithmetic problem, using a multiple linear regression with random effects for each player. *ROPE based on 85% statistical power. **ROPE based on 99% statistical power.

Pilot Study

We conducted a pre-registered (https://osf.io/udm8g/) pilot study (see Appendix for detailed results and full exclusion criteria). This study was designed to test the feasibility of the proposed design, not to test hypotheses. In conducting the pilot study, we underspecified exclusion criteria and deviated from the pre-specified pilot analysis plan. We consider all results from the pilot study to be exploratory.

The pilot study involved 48 participants. We excluded data from 1 participant that did not complete the study. This resulted in a final sample of 47 participants (23 female, 24 male).16 and 31 participants were assigned to the Competition and No-Competition treatments and 23 and 24 participants were assigned to the Effort and No-Effort

conditions, respectively. More participants were assigned to the No-Competition treatment because No-Competition participants need to participate first, so that participants in the Competition treatment have a Competitor to play against. By chance, we ended the pilot before more participants were recruited for the Competition treatment. The pilot design differed from the proposed design in one way: players were paid $0.25 cents per solution instead of $0.15 cents.

Participants in the Effort conditions spent more time (seconds) per grid (i.e. took longer to guess the majority color) than participants in the No-Effort conditions (95% HPDI: (11.81, 27.95), $\beta = 19.66$). This provides evidence that the Effort manipulation was successful. Participants in the Competition treatments had a larger log-odds of answering "yes" to an attention-check question about whether or not they competed against another player in the experiment (95% HPDI: (3.93, 22.67), $\beta = 11.92$). This provides evidence that participants in the Competition treatments were aware that they were competing against another player.

Compared to participants in the No-Competition treatments, participants in the Competition treatments revealed fewer tiles per grid (95% HPDI: (-7.46, -0.55), $\beta = -3.98$), did not have a lower log-odds of correctly guessing the majority color (95% HPDI: (-1.20, 0.03), $\beta = -0.60$), did not spend less time (seconds) per grid (95% HPDI: (-13.21, 3.92), $\beta = -5.02$), and made a larger number of guesses per minute (95% HPDI: (0.91, 4.12), $\beta = 2.50$). Compared to participants in the No-Competition x Effort treatment, participants in the Competition X Effort treatment solved more arithmetic problems per minute (95% HPDI: (0.55, 4.53), $\beta = 2.58$).

Compared to participants in the No-Effort conditions, participants in the Effort conditions did not reveal fewer tiles (95% HPDI: (-3.35, 2.12), β = -0.64), did not have a lower log-odds of correctly guessing the majority color (95% HPDI: (-0.62, 0.45), β = -0.07), spent more time (seconds) per grid (95% HPDI: (11.81, 27.95), β = 19.66), and made a smaller number of guesses per minute (95% HPDI: (-3.36, -0.63), β = -2.03).

There was no evidence for an interaction between Competition and Effort on number of tiles revealed (95% HPDI: (-1.51, 7.93), β = 3.26), log-odds of correctly guessing the majority color (95% HPDI: (-0.29, 1.53), β = 0.62), or time (seconds) spent per grid (95% HPDI: (-13.95, 8.67), β = -2.32). There was evidence for an interaction between Competition and Effort on the number of guesses made per minute (95% HPDI: (-4.49, -0.13), β = -2.28): in the No-Effort condition only, players in the Competition treatment make a larger number of guesses per minute than players in the No-Competition treatment. Again, we consider all of these results exploratory, as the pilot study was designed to test the feasibility of the proposed design, not to test hypotheses.

Results

We conducted the experiment according to the in-principle accepted Stage 1 protocol. Before data collection, we requested and received editorial approval for a minor deviation in the experimental procedure: instead of sessions where all participants are assigned to either the No-Competition or Competition treatment, we ran sessions such that participants could be assigned to either treatment within each session.

While data collection was ongoing, we requested and received editorial approval for another minor deviation: excluding data from all participants who informed the experimenter of technical difficulties during the experiment.

Every 4 sessions of data collection, we checked whether the HPDIs for all parameters fell entirely within or outside the pre-specified ROPEs for each hypothesis. This never occurred. As such, we collected data until we reached the pre-specified maximum sample size of 260 useable participants. In total, we collected data from 269 individuals. After applying the pre-specified exclusion criteria, our final sample size was 260 participants (6 participants did not complete the study and 2 participants experienced technical difficulties. 1 additional participant was excluded because every data point for their time-to-make-a-guess was above 5 standard deviations from the mean). The final sample of participants was composed of 130 females and 130 males (No-Effort condition: 65 females, 65 males; Effort condition: 65 females, 65 males).

Within individual participants, we excluded observations for which there was no data for at least one measured variable. We also excluded observations for time-to-make-a-guess and time-to-solve-arithmetic-problems that were more than 5 standard deviations larger than their respective means.

Below, we present analyses for quality checks and confirmatory predictions, using this final sample of 260 participants. All analyses (excluding exploratory analyses) use the exact statistical models specified earlier in this report and approved in Stage 1 review.

Quality Checks

Participants in the Effort conditions spent more time (seconds) per-click than participants in the No-Effort conditions (95% HPDI: (3.27, 3.99), $\beta = 3.64$). The effect falls entirely outside of the pre-specified ROPE of (-0.5, 0.5) and is in the predicted direction. This indicates that Effort Manipulation was successful.

Participants in the Competition treatments had greater log-odds of answering "yes" to the Competition Attention Check question than participants in the No-Competition treatments (95% HPDI: (2.85, 4.24), $\beta = 3.53$). The effect falls entirely outside of the pre-specified ROPE of (-0.8, 0.8) and is in the predicted direction. This indicates that participants in the Competition treatments were aware that they were competing against another player.

Confirmatory Analyses

We used Model 1 to test whether competition caused participants in the No-Effort condition to guess with smaller amounts of evidence (*H1a*) and whether competition caused a bigger reduction in the amount of evidence gathered in the No-Effort condition than in the Effort condition (*H3a*). Figure 3 plots the distribution of the raw data, alongside predicted means and 95% HPDI's from Model 1.

In the No-Effort condition, participants in the Competition treatment revealed fewer tiles per grid than participants in the No-Competition treatment (95% HPDI: (-5.03, -2.39), $\beta = -3.70$). The effect falls entirely outside of the pre-specified ROPE of (-1.22, 1.22) and is in the predicted direction. Based on our pre-specified criteria for evaluating hypotheses, this provides confirmatory evidence for *H1a*.

In the Effort condition, participants in the Competition treatment also revealed fewer tiles than participants in the No-Competition treatment (95% HPDI: (-4.47, -1.86), $\beta = -3.11$). There was no evidence for an interaction between Competition and Effort on number of tiles revealed (95% HPDI: (-1.23, 2.54), $\beta = 0.60$). The effect for the interaction does not fall entirely within or outside of the pre-specified ROPE of (-0.10, 0.10). Based on our pre-specified criteria for evaluating hypotheses, this does not provide conclusive evidence for *H3a* or for the null hypothesis.

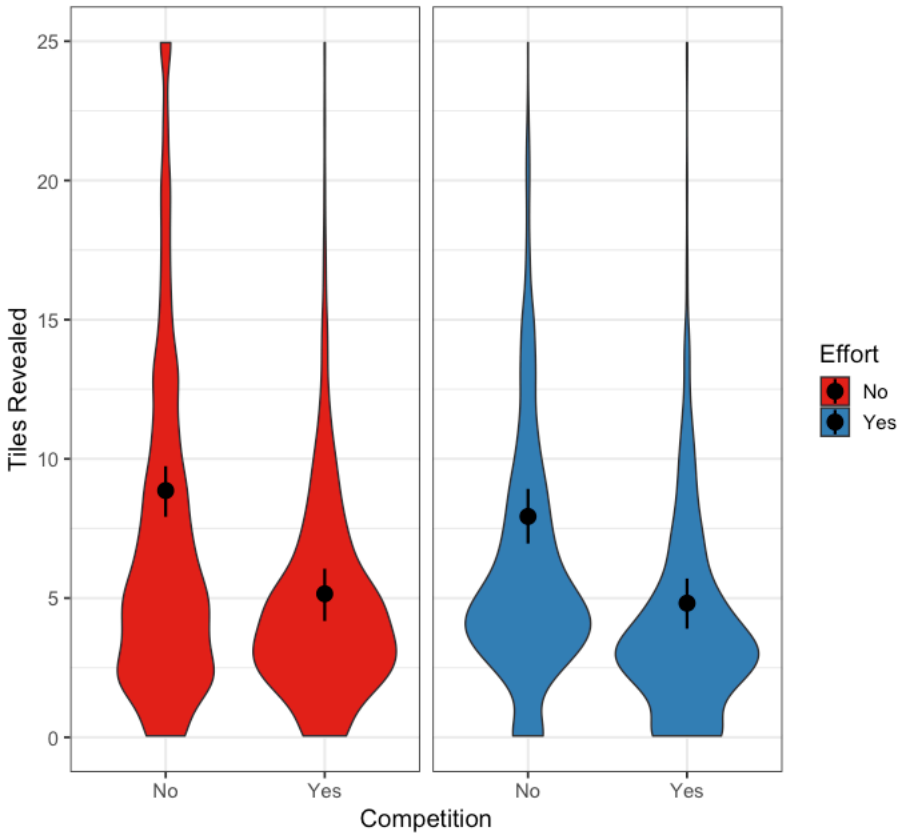


Figure 3| Tiles revealed. Participants revealed fewer tiles in the Competition, No-Effort treatment than in No-Competition, No-Effort treatment (95% HPDI: (-5.03, -2.39), $\beta = -3.70$). Participants also revealed fewer tiles in the Competition, Effort treatment than in No-Competition, Effort treatment (95% HPDI: (-4.47, -1.86), $\beta = -3.11$). There was no evidence that competition caused a larger

reduction in the number of tiles revealed in the No-Effort condition compared to the Effort condition (see main text).

Because *H1a* was confirmed, we used Model 2 to test whether competition caused participants in the No-Effort condition to have reduced accuracy (*H1b*). Because we did not find conclusive confirmatory evidence for *H3a*, we do not expect to find conclusive confirmatory evidence for *H3b*. Figure 4 plots the distribution of the raw data, alongside predicted means and 95% HPDI's from Model 2.

In the No-Effort condition, participants in the Competition treatment had a smaller probability of making a correct guess compared to participants in the No-Competition treatment (95% HPDI: (-0.11, -0.03), $\beta = -0.07$). In log odds, this effect is (95% HPDI: (-0.64, -0.20), $\beta = -0.42$), which falls entirely outside of the pre-specified ROPE of (-0.19, 0.19) and is in the predicted direction. Based on our pre-specified criteria for evaluating hypotheses, this provides confirmatory evidence for *H1b*.

In the Effort condition, participants in the Competition treatment also had a smaller probability of making a correct guess compared to participants in the No-Competition treatment (95% HPDI: (-0.11, -0.02), $\beta = -0.07$). There was no evidence for an interaction between Competition and Effort on the log-odds of making a correct guess (95% HPDI: (-0.30, 0.39), $\beta = 0.02$). The effect for the interaction does not fall entirely within or outside the pre-specified ROPE of (-0.09, 0.09). Based on our pre-specified criteria for evaluating hypotheses, this does not provide conclusive evidence for *H3b* or for the null hypothesis.

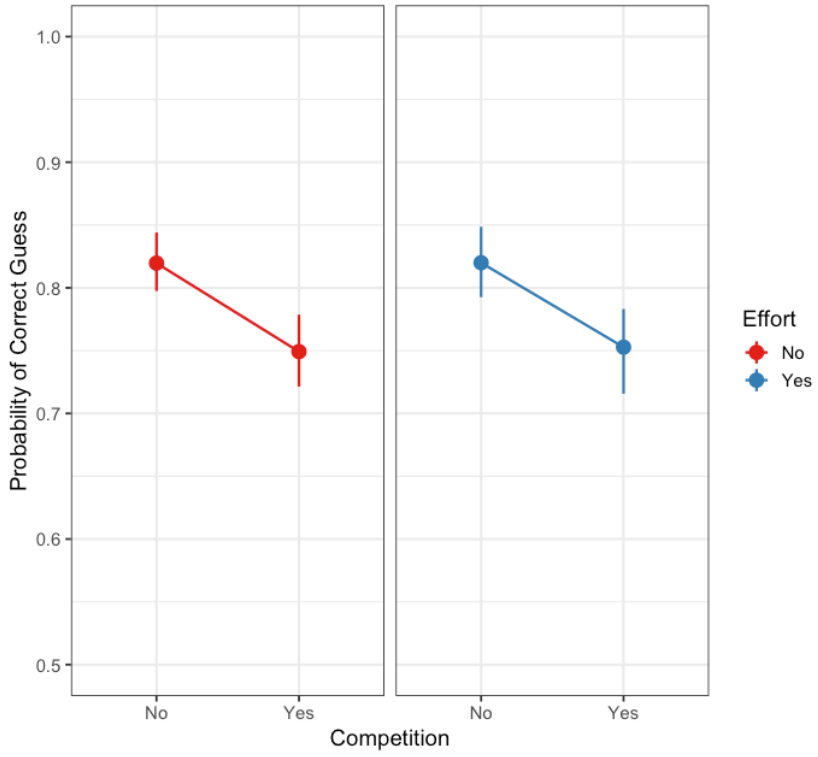


Figure 4| Accuracy. Probability of making a correct guess. Participants were less accurate in the Competition, No-Effort treatment than in No-Competition, No-Effort treatment (95% HPDI: (-0.11, -0.03), β = -0.07). Participants were also less accurate in the Competition, Effort treatment than in No-Competition, Effort treatment (95% HPDI: (-0.11, -0.02), β = -0.07). There was no evidence that competition caused a larger reduction in accuracy in the No-Effort condition compared to the Effort condition (see main text).

We used Model 3 to test whether competition increased effort: participants in the Competition-Effort treatment should faster to accurately solve one arithmetic problem than participants in the No-Competition, Effort treatment (*H2*). Figure 5 plots the distribution of the raw data, alongside predicted means and 95% HPDI's from Model 3.

Participants in the Competition-Effort treatment were not faster to accurately solve one arithmetic problem than participants in the No-Competition, Effort treatment (95% HPDI: (-0.40, 0.35), β = -0.02). The effect does not fall entirely within or outside of

the pre-specified ROPE of (-0.33, 0.33). Based on our pre-specified criteria for evaluating hypotheses, this does not provide conclusive evidence for *H2* or for the null hypothesis.

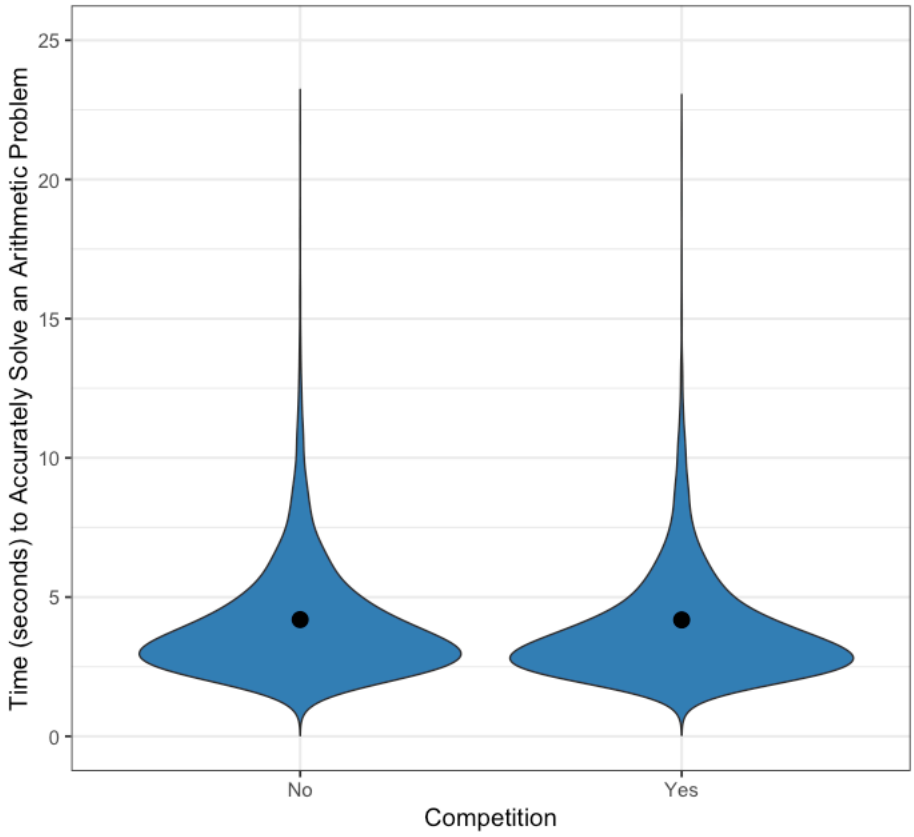


Figure 5| Time (seconds) to accurately solve one arithmetic problem. Competition did not increase participant effort: participants in the Competition-Effort treatment did not have a faster time to solve one arithmetic problem than participants in the No-Competition, Effort treatment (95% HPDI: (-0.40, 0.35), β = -0.02).

Exploratory Analyses

We tested the relationship between the ratio of tiles on a given grid (i.e. effect size) and the number of tiles revealed by participants by modifying Model 1 to include interactions between effect size and competition/effort, and control for guess number. Across all treatments and conditions, players generally revealed fewer tiles as effect size

increased (Figure 6). There was exploratory evidence for a negative effect of effect size on number of tiles revealed in the No-Competition, No-Effort treatment (95% HPDI: (-0.50, -0.65), $\beta = -0.57$). There was also exploratory evidence for a positive interaction between the Competition treatment and effect size on number of tiles revealed (95% HPDI: (0.20, 0.39), $\beta = 0.29$), and a positive interaction between the Effort condition and effect size on number of tiles revealed (95% HPDI: (0.09, 0.30), $\beta = 0.19$).

Within treatments, in the No-Effort, No-Competition treatment, participants revealed fewer tiles for large (95% HPDI (7.19, 8.69), mean = 7.94) compared to small (95% HPDI (8.76, 10.24), mean = 9.48) effect sizes. In the No-Effort, Competition treatment, participants did not reveal fewer tiles for large (95% HPDI (4.00, 5.60), mean = 4.83) compared to small (95% HPDI (4.8, 6.40), mean = 5.59) effect sizes. In the Effort, No-Competition treatment, participants did not reveal fewer tiles for large (95% HPDI (6.63, 8.24), mean = 7.48) compared to small (95% HPDI (7.69, 9.28), mean = 8.49) effect sizes. In the Effort, Competition treatment, participants did not reveal fewer tiles for large (95% HPDI (3.6, 5.21), mean = 4.37) compared to small (95% HPDI (3.79, 5.42), mean = 4.6) effect sizes.

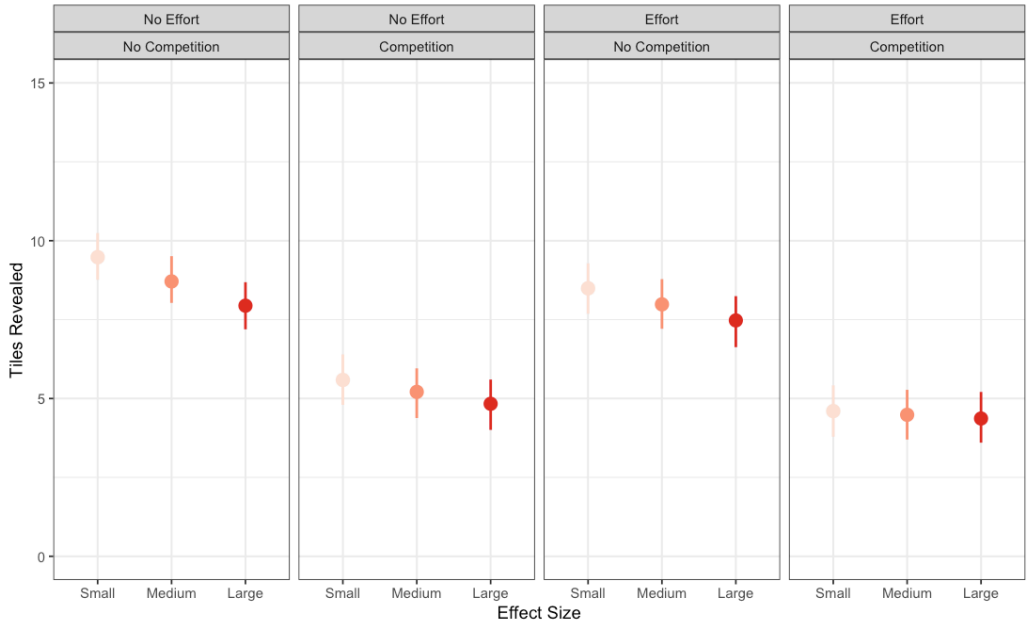


Figure 6| Tiles revealed as a function of effect size. Participants generally revealed fewer tiles on grids with larger effect sizes. There was a positive interaction between the Competition treatment and effect size, and between the Effort condition and effect size, on number of tiles revealed (see main text).

We also tested the relationship between effect size and accuracy by modifying Model 2 to include interactions between effect size and competition/effort, and control for guess number. Across all treatments and conditions, players had higher accuracy as effect size increased (Figure 7). There was exploratory evidence for a positive effect of effect size on log-odds of making a correct guess in the No-Competition, No-Effort treatment (95% HPDI: (0.59, 0.72), $\beta = 0.66$). There was also exploratory evidence for a negative interaction between the Competition treatment and effect size on log-odds of making a correct guess (95% HPDI: (-0.23, -0.06), $\beta = -0.14$), and a positive interaction between the Effort condition and effect size on log-odds of making a correct guess (95% HPDI: (0.06, 0.26), $\beta = 0.16$).

Within treatments, in the No-Effort, No-Competition treatment, participants had a larger probability of making a correct guess for large (95% HPDI (0.91, 0.93), mean = 0.92) compared to medium (95% HPDI (0.81, 0.85), mean = 0.83) and small (95% HPDI (0.64, 0.70), mean = 0.67) effect sizes. In the No-Effort, Competition treatment, participants had a larger probability of making a correct guess for large (95% HPDI (0.84, 0.88), mean = 0.86) compared to medium (95% HPDI (0.73, 0.78), mean = 0.76) and small (95% HPDI (0.57, 0.64), mean = 0.61) effect sizes. In the Effort, No-Competition treatment, participants had a larger probability of making a correct guess for large (95% HPDI (0.91, 0.94), mean = 0.93) compared to medium (95% HPDI (0.79, 0.84), mean = 0.81) and small (95% HPDI (0.55, 0.63), mean = 0.59) effect sizes. In the Effort, Competition treatment, participants had a larger probability of making a correct guess for large (95% HPDI (0.85, 0.89), mean = 0.87) compared to medium (95% HPDI (0.70, 0.76), mean = 0.73) and small (95% HPDI (0.48, 0.57), mean = 0.53) effect sizes.

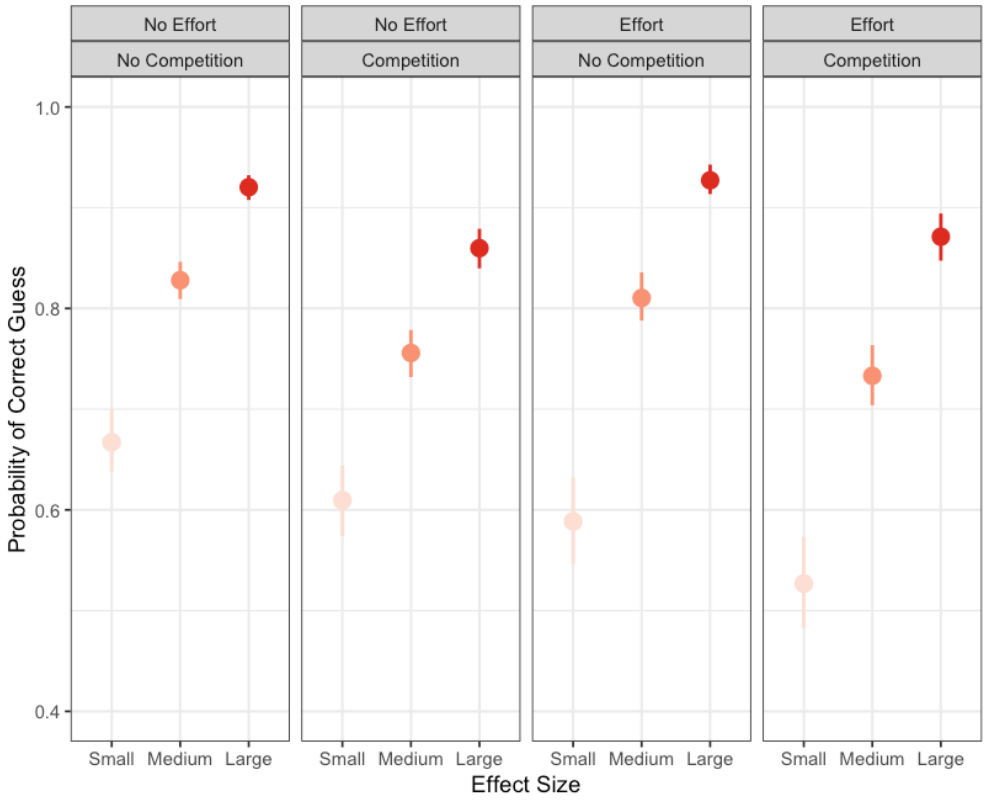


Figure 7| Accuracy as a function of effect size. Participants had a higher probability of making a correct guess for larger effect sizes. There was a negative interaction between the Competition treatment and effect size, and a positive interaction between the Effort condition and effect size, on accuracy (see main text).

Summary

We developed a laboratory experiment to test how competition for priority affects information sampling in a game that parallels scientific investigation. The pilot study indicated that the effort and competition manipulations were successful and produced no indication of floor or ceiling effects for any dependent measure. This study was then approved for in-principle acceptance as a Registered Report at *Royal Society Open Science*. Following our pre-specified experimental protocol, we collected data from 260 students at Arizona State University. Pre-specified quality checks revealed that the

experimental manipulations were successful. Confirmatory analyses revealed that competition caused participants to make guesses with less information, thereby reducing their accuracy. Confirmatory analyses did not provide conclusive evidence for an interaction between competition and effort, and did not provide evidence that competition caused participants to increase effort (i.e. time to accurately solve an arithmetic problem). The 95% HPDI around the effect of competition on effort (-0.40, 0.35) indicates that any existing effect of competition on time to solve arithmetic problems is likely to be small. This study provides experimental evidence that competition for priority can cause individuals to make guesses based on less information. The design is merely one instantiation of priority races and the results should not be interpreted as providing definitive evidence that incentivizing priority harms the scientific process. Whether incentivizing novel findings harms the scientific process remains an open question, and will depend on the extent to which these results can be replicated and can generalize across different experimental instantiations of competition for priority.

Constraints on Generality (COG)

We provide a statement of the *Constraints on Generality (COG)* of our experiment (Simons, Shoda, & Lindsay, 2017b). *Participants.* We have no reason to believe that the effect of competition on information sampling depends on characteristics of participants. The effects should replicate when scientists participate in this experiment. *Materials.* The effects should not depend on the specific colors of the two different tiles or the number of different underlying effect sizes. We make no claims as to whether the results depend on other characteristics of the materials used in this study. *Procedures.*

Participants should pass a tutorial and comprehension check before starting the study. Participants should not be able to see the performance of other simultaneous participants in the study. We make no claims as to whether the results generalize to situations in which two participants directly compete against one another and can dynamically respond to each other's behavior. *Historical/Temporal Specificity.* We have no reason to believe that the results depend on characteristics of historical or temporal specificity.

Ethics

Permission to perform this study was granted by the Arizona State University Institutional Review Board (IRB), code: STUDY00007691. All participants provided informed consent.

Data Accessibility

All data, materials (including experimental protocols) and code are openly available at the Open Science Framework (https://osf.io/udm8g/). The pilot pre-registration can be found at the Open Science Framework (https://osf.io/udm8g/).

CHAPTER 3

COMPETITION FOR PRIORITY AND THE NATURAL SELECTION OF UNDERPOWERED RESEARCH

Abstract

It is becoming increasingly clear that many published findings do not replicate (i.e. the "reproducibility crisis"). As a consequence, scholars in diverse disciplines are interested in understand how different factors affect the reliability of science and in evaluating potential solutions. Currently, incentive structures are thought to be one key determinant of the reliability of research that scientists conduct. Here we develop an evolutionary agent-based model to test the effect of incentives for priority of discovery on the reliability of scientific findings. In this model, scientists investigate a phenomenon and compete to be first to obtain a statistically significant result. Scientists can increase statistical power by using larger samples, but this takes more time and so increases their risk of being "scooped". We find that competition for priority causes populations of scientists' practices to culturally evolve towards lower sample sizes and, in turn, lower statistical power. This mirrors the results of a previous model about the cultural evolution of bad science. However, we also find that two factors attenuate the negative effects of competition: increased time costs associated with setting up a single study and increased payoffs to publication of secondary (i.e. scooped) results. Startup costs lower the relative payoff to individuals who pursue a "quantity" strategy by conducting many low-quality studies. Payoffs for "scooped" results allow individuals who conduct few high-quality studies to obtain benefits even though they are frequently scooped. We discuss the

implications of these findings for preventing low-quality research and use them to evaluate the effectiveness of proposed scientific reforms (e.g. registered reports).

Introduction

In science, priority matters. The norms of science have rewarded researchers for being first to make discoveries for at least several hundred years (Merton, 1957). Such rewards take various forms, including financial prizes for discoveries (e.g. the Nobel prize), increased probability of obtaining professional positions or speaking engagements, and increased probability of publishing one's results in a high-impact academic journal (Fang & Casadevall, 2015; Makel et al., 2012; Nosek et al., 2012). A consequence of this norm is that scientists are strongly incentivized to compete over priority.

In recent years, there has been growing recognition that some scientific incentives contribute to erroneous published findings (Ioannidis, 2014; Munafò et al., 2017; Nosek et al., 2012), by favoring positive and novel findings, evaluating scientists based on quantity rather than quality of publications, and disincentivizing data sharing and transparent research (Begley & Ioannidis, 2015; Higginson & Munafò, 2016; Munafò et al., 2017; Nissen et al., 2016; Smaldino & McElreath, 2016). Among many such incentives, scholars have been especially concerned about the harmful effects of competition on scientific outcomes (Alberts et al., 2014; Anderson et al., 2007; Benedictus et al., 2016; Fang & Casadevall, 2015; Geman & Geman, 2016; Nosek et al., 2012; Rawat & Meena, 2014; Sarewitz, 2016; Smaldino & McElreath, 2016). In competitive incentive structures, individuals can expend finite resources (e.g. time, money) to increase their probability of receiving a payoff, and one individual's success

reduces the probability that others will succeed (Dechenaux, Kovenock, & Sheremeta, 2015). By this definition, much of science is competitive; scientists seek to gain access to limited resources such as funding and faculty positions, and their success implies the failure of others. Being first to report a given result helps scientists gain access to funding and faculty positions, which in turn causes competition over priority.

Given its role as a major scientific incentive, how does rewarding priority of discovery affect the scientific process? It is possible that rewarding priority benefits science. For instance, scientists may be incentivized to quickly solve problems and share their findings with the scientific community (Merton, 1957). Scientists may also efficiently distribute themselves among multiple scientific problems, because each scientist benefits from working on a problem with as few competitors as possible (Bergstrom et al., 2016; Strevens, 2003). The prospect of losing out in a competitive system may also increase the individual effort, task performance, and innovation of scientists, relative to a system in which individuals are rewarded for each unit of output regardless of order. This could improve the quality of science that is carried out (Balietti et al., 2016; Dechenaux et al., 2015; Gneezy et al., 2003).

However, rewarding priority has plausible downsides. Scientists may rush their work in an effort to avoid being scooped (Yong, 2018a). This might reduce the quality of their research by increasing their probability of making mistakes or by reducing the amount of information that they share with the scientific community. Robert Merton described Charles Darwin's irritation with how rewarding priority rewarded rushed, low-quality work, writing 'In biology, it is the long-standing practice to append the name of

the first describer to the name of a species, a custom which greatly agitated Darwin since, as he saw it, this put "a premium on hasty and careless work" as the "species mongers" among naturalists try to achieve an easy immortality by "miserably describing a species in two or three lines"' (Merton, 1957, p. 644). Scientists themselves are aware of how competition for priority incentivizes rushed work. In a set of focus-group discussions with 51 researchers, one early-career researcher described this effect: "There's a fine line with actually having enough ... data to support your idea, and then going that extra half-meter to really send it home. You don't have that sort of time, because if you don't get it published in a timely fashion, someone else will—without that data." (Anderson et al., 2007, p. 458).

Here we focus on the latter hypothesis: competition for priority incentivizes rushed, low-quality research. To evaluate the plausibility of this hypothesis, we develop an evolutionary agent-based model that tests the effect of incentive structures that reward priority of discovery on the scientific process. In our model, researchers simultaneously compete to be first to publish a significant result on a given problem. Researchers can increase their statistical power by increasing their sample size. However, increasing sample size costs time, which leaves researchers vulnerable to being scooped by competitors.

Our model supports the general hypothesis that competition can reduce the quality of individual studies: the practices of scientists who compete for priority culturally evolve towards smaller equilibrium sample sizes and lower statistical power than those for individual scientists who are not in competition. We also propose a mechanism that

allows populations to maintain higher sample sizes and statistical power at equilibrium: increased costs to starting investigations. This finding is robust to several extensions of the initial model (see below and Appendix). This suggests that competition for priority can cause researchers to pursue lower-quality investigations, but that its effect on science depends on the cost of starting new investigations. It also suggests that the time-cost inherent to some current reforms (e.g. pre-registration; registered reports) may be a feature rather than a bug, because it de-incentivizes researchers from conducting large numbers of underpowered studies.

Model

Consider a population of n scientists. Each scientist is characterized by a single (positive, discrete) parameter, s, that corresponds to their sample size when they conduct research. On any given question, a scientist's statistical power (*pwr)* is a function of s, the false-positive rate (α, i.e., the significance threshold), and the size of the effect being studied (e). A scientist's *pwr* is calculated using a two-sample t-test, by passing these parameters to the pwr.t.test() function in the 'pwr' package in R (Champely et al., 2018; R Core Team, 2017). In effect, this assumes that all research is of the form where scientists collect s data points from each of two populations and then test for a difference between the two.

In our model, e values are drawn from an exponential distribution characterized by the rate parameter λ (distinct from the aforementioned *rate* of significant results) and rounded to the nearest 0.1. Each scientist's career lasts for T time-steps. Once their career

has started, scientists collect data until they reach their desired sample size as dictated by their respective *s* value. The number of time steps required to do this (*t*) is:

$$t = s * c_s + c$$

Once a scientist has completed data collection, they perform a significance test, with probability *pwr* of obtaining a statistically-significant result. Every time a scientist obtains a statistically-significant result, they gain 1 point and move on to the next problem, starting a new study from scratch. If they are competing against other scientists, then all other scientists also move on to the next problem. In this way only the first player to obtain a statistically-significant finding receives any points, because being scooped prompts scientists to abandon their current work. However, if a scientist tests their data and fails to find a significant result, their competitor continues as usual, while the focal scientist starts a new study from scratch. If both scientists conclude a study and obtain a statistically-significant result at the same time, each scientist obtains ½ point.

We allow the phenotypes of scientists to evolve across generations: scientists who acquire more points within a generation have a higher probability of being represented in the next generation. This evolutionary component of our model corresponds to the assumption that scientists who are more successful (e.g. have more publications) are more likely to pass on their characteristics (e.g. the sample size used in their studies) to the subsequent generation of scientists. For example, younger scientists may preferentially imitate the behaviors of successful, well-established scientists (i.e. payoff-biased social learning, (Richard McElreath et al., 2008)). Alternatively, scientists who acquire more publications may be more likely to remain in academia, and will thus be

disproportionately available as cultural models for other scientists (Brischoux & Angelier, 2015; Smaldino & McElreath, 2016; van Dijk, Manor, & Carey, 2014). Our model is agnostic to the specific mechanism of cultural transmission by which the characteristics of successful scientists become disproportionately represented in subsequent generations.

The cultural-fitness of each scientist is proportional to their total number of points. After all rounds of competition are complete, individuals reproduce according to the Wright-Fisher model, where repeated sampling with replacement from the parent generation, weighted by fitness, generates n offspring. The sample size (s) of offspring is a rounded value drawn from a normal distribution with mean s_{parent} and standard deviation 2. Values of $s < 2$ are set to 2, because two-sample t-tests require at least 2 samples in each group. The sample sizes of the initial population of scientists are generated by rounding n random numbers drawn from a uniform distribution ranging from 2 to 1000.

Individual scientists (i.e. no competition).

Assume first that α is constant across all studies, that $\lambda = 3$ (i.e. e values are drawn from an exponential distribution with a rate parameter = 3), and that there is no competition (i.e. scientists' payoffs are determined entirely by their rate of statistically-significant results). Given these parameters, we can approximate the payoff-maximizing sample size for individual scientists by evolving a population of $n = 100$ scientists across many generations do determine the equilibrium sample size. Figures 1a and 1b plot

equilibrium sample size and statistical power (respectively) for individual scientists, as a function of varying startup costs (s_c), after 200 generations of evolution.

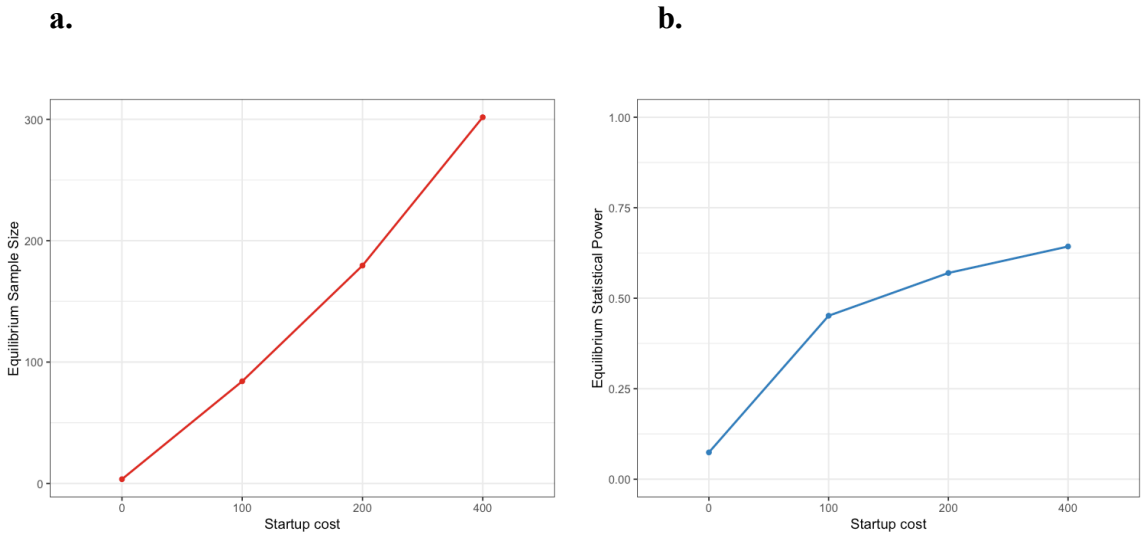


Figure 1 | Equilibrium (a) sample size and (b) statistical power for individual scientists as a function of startup cost (200 generations, 50 repeats). Parameter values are: n=100, α=0.05, $\lambda = 3$, r=5, T=5000 and c_s=1. Larger startup costs lead to larger sample sizes and statistical power at equilibrium.

When there are no startup costs, populations of scientists evolve towards very small sample sizes. As startup costs increase, equilibrium sample size increases. Scientists who conduct studies with small sample sizes have low statistical power which means that their probability of obtaining a statistically-significant result in a given study is low. Instead, their success depends on performing many studies as quickly as possible. This is most profitable when startup costs are low because scientists can perform multiple successive studies quickly. When the goal is to obtain at least one statistically-significant finding, running many small, underpowered studies can be a more efficient strategy than running one larger, well-powered study. This result is consistent with prior simulations of

strategies for chasing statistical significance in science (Bakker, van Dijk, & Wicherts, 2012).

However, large startup costs de-incentivize researchers from pursuing such a "quantity" strategy because they place a time cost on the scientist every time they start (or restart) a study. As startup costs increase, researchers obtain higher payoffs by investing more in each individual study – the additional time required to collect a larger dataset being compensated by the increased probability of finding a significant result. To illustrate, consider a case where $e = 0.1$. A scientist with a sample size of 100 has statistical power of 0.11, whereas a scientist with sample size $s = 300$ has statistical power = 0.23. When there are no startup costs, the $s = 100$ scientist can run 3 studies during the time that the sample size $s = 300$ scientist can run 1. The former's probability of detecting at least 1 statistically-significant effect within 300 time periods is 1 - 0.89^3 = .295, while the latter's is 0.23. In this case, the $s = 100$ scientist is likely to win. Now consider a case where the startup cost is 300. It will take the $s = 100$ scientist 400 time-periods (100 + 300) to conduct one study with statistical power = 0.11, and it will take the $s = 300$ scientist 600 time-periods (300 + 300) to conduct one study with statistical power = 0.23. The former's probability of obtaining at least 1 statistically-significant effect after 800-time periods is 1 – 0.89^2 = 0.21, while the latter's probability of detecting a significant effect after just 600 time-periods is 0.23. Thus, higher startup costs decrease the relative payoff of small sample-size scientists compared to large sample-size scientists.

Extension 1: Multiple Competitors and Sample Sizes Drawn from Different Exponential Distributions

Thus far, we have assumed that effect sizes were drawn from an exponential distribution with $\lambda = 3$ and that $n = 1$ (i.e. there was no competition). Below, we modify these assumptions.

As a sensitivity check, we explore the effect of drawing *e* valued from a range of exponential distributions, that vary in their λ parameter. We also incorporate competition: scientists compete against other scientists to be first to detect statistically-significant results (see "Model" description above for details). To test the effect of increasing competition on equilibrium sample size, we vary the number of competitors across simulations: scientists compete against each other in groups of 2, 4, or 8. As before, only the first scientist to obtain a statistically-significant result obtains a payoff, and all scientists simultaneously move on to the next problem upon being scooped. Each group is composed of individuals randomly-selected (without replacement) from the population of *n* scientists, where $n = 100$ (for groups of 2 and 4) or $n = 96$ (for groups of 8). Competition occurs locally (i.e. within each group). However, fitness is determined globally (i.e. a scientist's proportion of the total number of points of all scientists in the population) and reproduction occurs in the same way as before (see "Model" description above).

Figures 2 illustrates effect of competition on sample size by plotting equilibrium sample as a function of varying startup costs (s_c) and number of competitors, when $\lambda = 3$. The appendix for Chapter 3 contains an alternative visualization of this effect. Given any

level of competition and non-zero startup costs, equilibrium sample size is substantially lower than for individual researchers. This occurs because more competitors increase the probability that any given scientist will be scooped, which favors smaller sample sizes. To illustrate, consider a case where $e = 0.2$, there are no startup costs, and where there are two competitors with sample sizes 50 and 150, respectively. The scientist with a sample size of 50 has statistical power of 0.17, whereas the scientist with a sample size of 150 has statistical power of 0.41. The $s = 50$ scientist can run 2 studies (i.e. at time periods 50 and 100) before the $s = 150$ scientist is even able to run 1. The former's probability of detecting at least 1 statistically-significant effect before the latter is even able to run 1 is 1 - 0.83^2 = .31. In this case, the $s = 100$ scientist has approximately a 30% chance of being scooped before being able to run even 1 study. Now consider a case where the $s = 100$ scientist faces 7 other competitors, all of whom have $s = 50$. There is now a 1 - 0.69^7 = 0.93 probability that at least one other scientist obtains at least 1 statistically-significant result before the $s = 100$ scientist is even able to run 1 study.

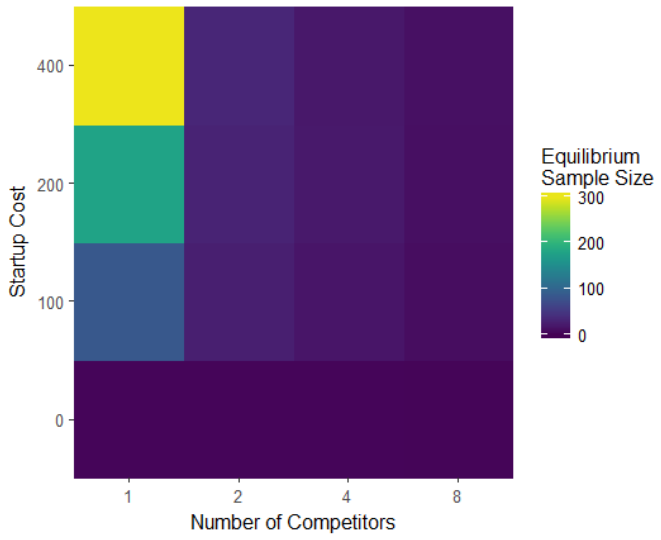


Figure 2 | Equilibrium sample size for individual scientists compared to varying numbers of competitors, as a function of startup cost (200 generations, 50 repeats). Parameter values are: n=100, α=0.05, $\lambda = 3$, r=5, T=5000 and c_s=1. For any number of competitors (i.e. 2, 4, 8), equilibrium sample size is lower than that of individual scientists (i.e. competitors = 1). As the number of competitors increases, equilibrium sample size decreases, because more competitors increase the probability that any given researcher will be scooped. As startup costs increase, equilibrium sample size increases, for the same reasons as in Figure 1. See Figure 5S in Appendix for an alternative visualization of Figure 2.

Figures 3 and 4 plot equilibrium sample size and statistical power, respectively, for exponential distributions that vary in their λ parameter. To better illustrate the effect of increasing competition and startup costs, these figures only display equilibrium sample size given some level of competition (i.e. 2, 4, or 8 competitors). Equilibrium sample size for individual researchers (i.e. competitor = 1) is always substantially higher than equilibrium sample size given any amount of competition. Figures 3 and 4 demonstrate that the pattern in Figure 2 generalizes to a range of effect-size distributions. More competitors lead to smaller equilibrium sample sizes, while higher startup costs lead to larger equilibrium sample sizes. Equilibrium sample size is largest for intermediate values of λ and smallest for small and large values of λ. When λ is large, effect sizes are

often zero, which disfavors investment in large samples. When λ is intermediate, scientists with larger samples are better able to detect the increased number of small and medium-sized effects. When λ is small, effect sizes are enormous, and scientists can detect them even with small sample sizes. However, equilibrium statistical power is highest for small values of λ: effect sizes are often large, and even scientists with small sample sizes can often obtain substantial statistical power.

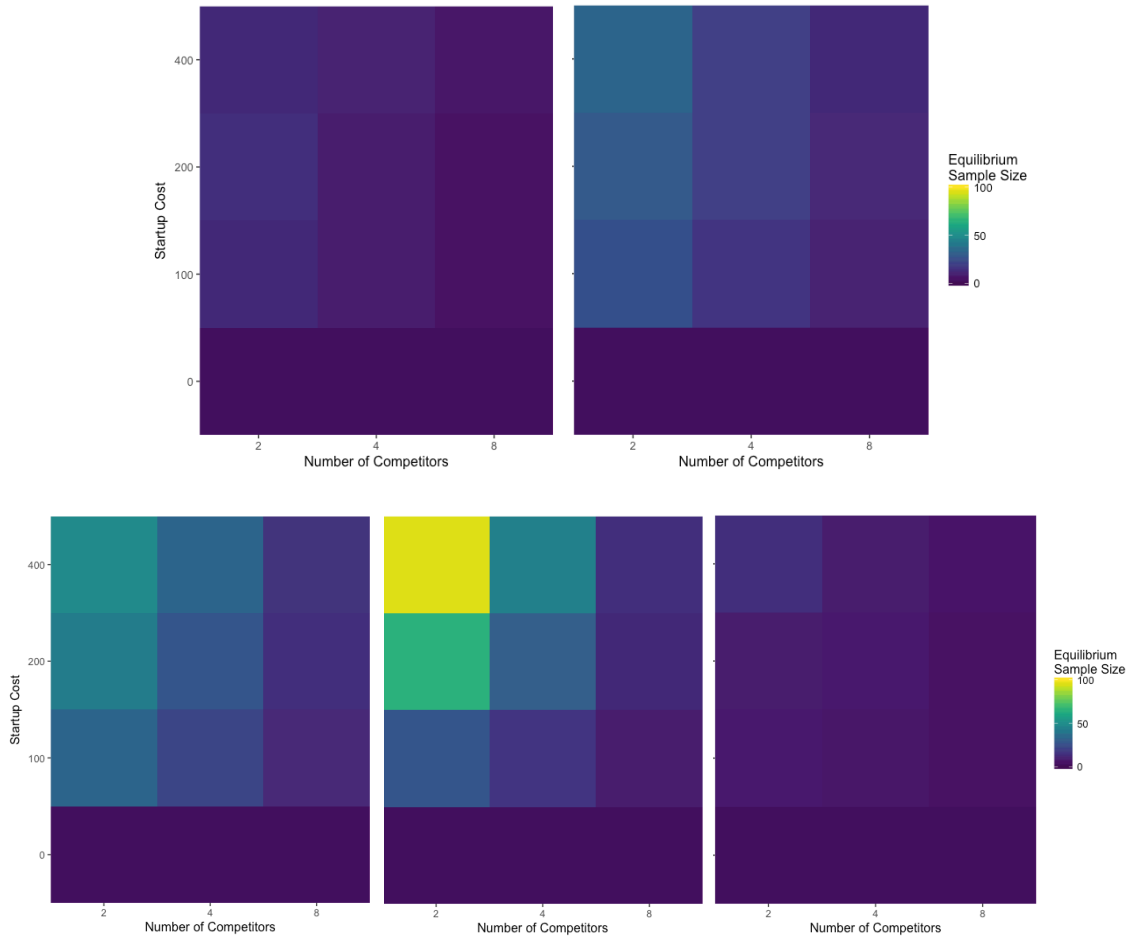


Figure 3 | Equilibrium sample size as a function of number of competitors, startup cost, and exponential-distribution rate (λ) parameter (200 generations, 50 repeats). Parameter values are: n=100, α=0.05, r=5, T=5000 and c_s=1. We explored the effect of running the simulation with five levels of λ (top row, left to right: 1, 3; bottom row, left to right: 5, 10, 50). As number of competitors increases, equilibrium sample size decreases, because more competitors increase the probability that any given scientist will be scooped. As startup costs increase, equilibrium sample size increases, for the same reasons as in other simulations. Equilibrium sample size is largest for intermediate values of the rate parameter and smallest for small and large values. When λ is large, effect sizes are often zero, which disfavors investing in large samples. When λ is intermediate, scientists with larger samples are better able to detect the increased number of medium-sized effects. When λ is small, effect sizes are enormous, and scientists can detect them even with small sample sizes.

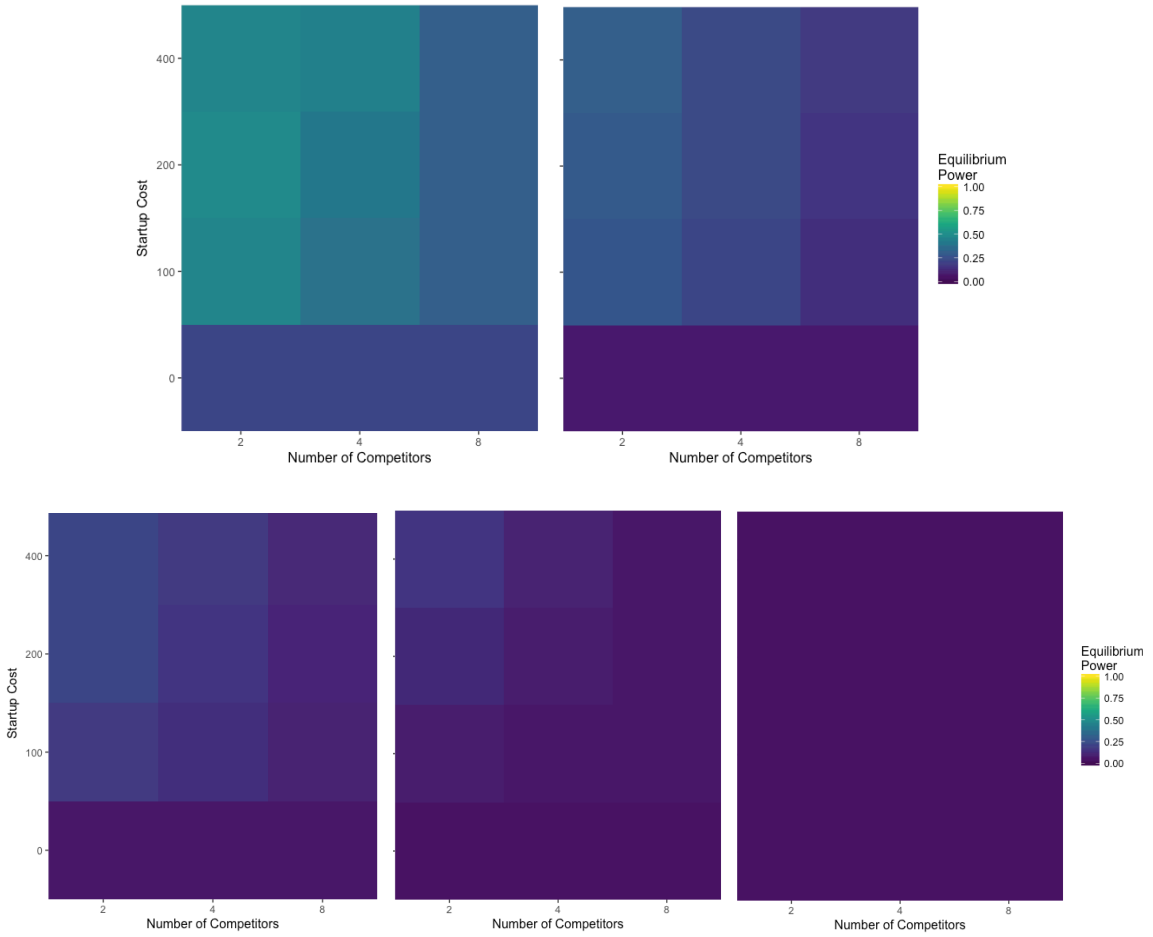
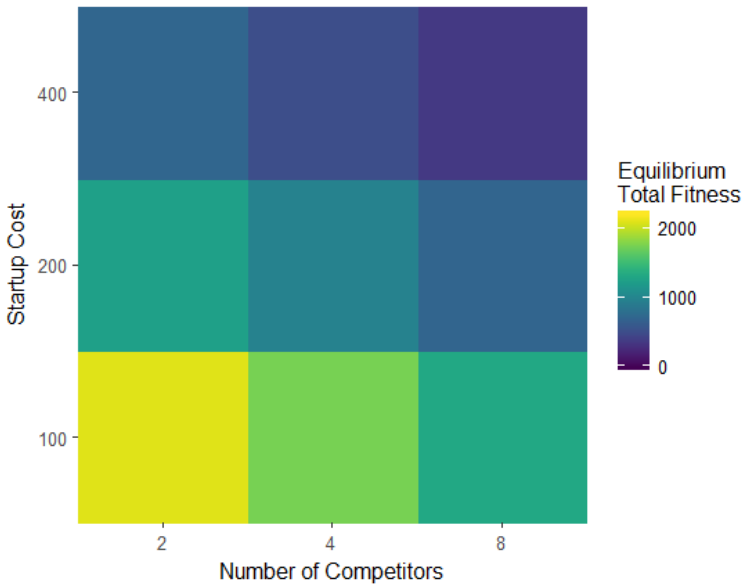


Figure 4 | Statistical power as a function of number of competitors, startup cost, and exponential-distribution rate (λ) parameter (200 generations, 50 repeats). Parameter values are: n=100, α=0.05, r=5, T=5000 and c_s=1. We calculated the statistical power at equilibrium, given five levels of λ (top row: 1, 3; bottom row: 5, 10, 50). When effects are large (e.g. λ = 1) statistical power is high, even though equilibrium sample size is low. This is because scientists do not need large samples to obtain high statistical power, given large effect sizes. When effects are small (and often 0, e.g. λ = 50), average statistical power is low, even though equilibrium sample size is often substantial. This is due to two factors. For small effects, statistical power increases very slowly with increasing sample size, meaning that even large samples do not provide much statistical power. Further, when effects are 0, statistical power is fixed at the false-positive rate and does not increase with increasing sample size. As startup costs increase, statistical power increases. This occurs because startup costs increase equilibrium sample size, and statistical power is a monotonically increasing function of sample size.

Figure 5 plots equilibrium total fitness (i.e. the combined number of false positive and true positive results) as a function of startup cost and number of competitors. This

provides a measure of the total number of studies completed by all scientists. Higher startup costs lead to fewer total positive results. This occurs because scientists spend much time waiting (instead of sampling) and are able to conduct fewer studies. A larger number of competitors also decreases the total number of positive results. This occurs because more competition causes many scientists to compete over the same research questions (i.e. effects). Although this increases the speed with which a statistically-significant finding is found for a given research question, it decreases the total number of research questions investigated by the population. For instance, when groups consist of 2 competitors and the population consists of $n = 100$ scientists, 50 groups of 2 scientists each collect data on 50 unique sequences of effects. When groups consist of 4 competitors, 25 groups of 4 scientists each collect data on 25 unique sequences of effects. The decrease in the total number of questions investigated by the population of scientists results in fewer total positive results with increased competition.



.

Figure 5 | Total fitness (i.e. number of positive results) as a function of number of competitors and startup cost (200 generations, 50 repeats). Parameter values are: n=100, α=0.05, $\lambda = 3$, r=5, T=5000 and c_s=1. We explored the total number of positive (i.e. statistically-significant results) as a function of the number of competitors and startup cost. As number of competitors increases, the total number of positive results decreases. This occurs because many researchers compete over the same research questions (i.e. effects). Although this increases the speed with which a statistically-significant finding is found for a given research question, it decreases the total number of research questions investigated by the population. As startup costs increase, the total number of positive results decreases. This occurs because researchers conduct fewer studies when startup costs are high, which means that they have a slower rate of obtaining statistically-significant results per unit time.

Extension 2: Adding benefits for secondary publications

So far, our model has assumed that scientists who get scooped receive no payoff: they simply abandon their current investigation and move on to the subsequent one. Below, we modify this assumption, allowing for payoffs to secondary publication. This corresponds to the assumption that scientists who are scooped are still sometimes able to publish their findings or receive other forms of recognition for their research, which results in some non-zero payoff. For instance, the academic journals eLife and PLOS

Biology have recently begun offering "scoop protection" (i.e. allowing researchers to publish findings identical to those already published in the same journal) in attempts to reduce the disproportionate payoffs to scientists who publish first (Editors, 2018; Marder, 2017; Yong, 2018a). Extending the model to allow for benefits to secondary publication allows us to evaluate the effects of such policy changes on research quality.

In this extension, scientists who are scooped stop sampling and conduct a significance test. The significance test has the statistical power (*pwr*) associated with the number of participants that the scientist had gathered at the time of being scooped. With probability *pwr*, the scientist obtains a statistically-significant result and obtains the secondary benefit, b_2. With probability $1 - pwr$, they obtain a null result and receive no payoff. Scientists then move on to the subsequent research question (i.e. all competitors remain synchronized).

Figure 6 plots equilibrium sample size (Log_{10}) as a function of startup cost and number of competitors, for varying levels of b_2. Given any non-zero level of startup costs, benefits for secondary publication increase equilibrium sample size. This occurs because benefits to secondary publication allow scientists who are most likely to get "scooped" to receive a non-zero payoff. This reduces the difference in relative payoffs between scooped scientists and those who are first to obtain a statistically-significant finding. As long as scientists with large sample-sizes are more likely to be scooped than those with small sample-sizes, increasing the rewards to scooped scientists should increase equilibrium sample size.

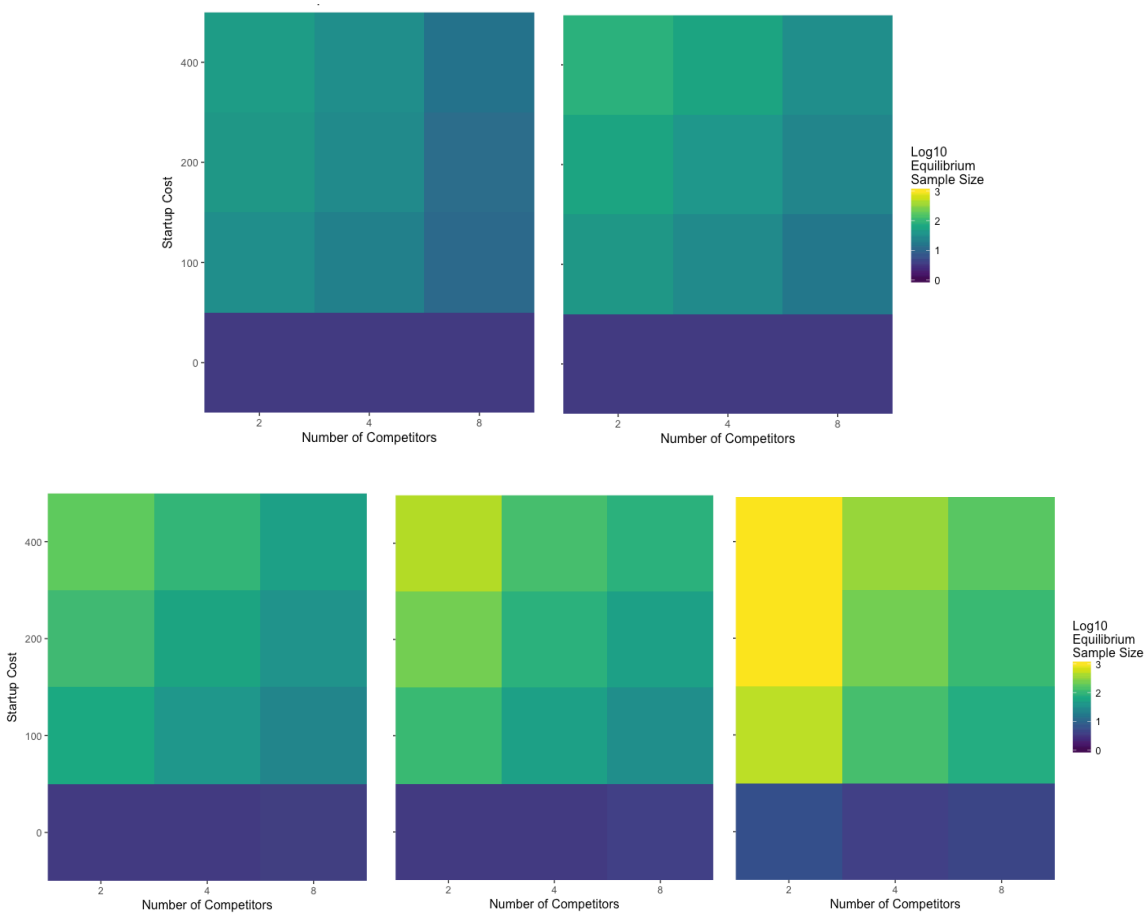


Figure 6 | Equilibrium sample size (Log_{10}) as a function of number of competitors and startup cost, for various levels of benefit to secondary publication, b_2 (200 generations, 50 repeats). Parameter values are: n=100, α=0.05, $\lambda = 5$, r=5, T=5000 and c_s=1. We explored the effect of running the simulation with five levels of benefits to secondary publication (top row, left to right: 0, 0.33; bottom row, left to right: 0.66, 1, 2). Once at least one scientist has published a statistically-significant result, all other scientists stop collecting data and test for statistical significance, where their statistical power is determined by their sample size at the time of being scooped. Any of those scientists who obtain a statistically-significant result receives the secondary benefit. As number of competitors increases, equilibrium sample size decreases, because more competitors increase the probability that any given researcher will get scooped. As startup costs increase, equilibrium sample size increases. Larger benefits to secondary publication increase equilibrium sample size. Secondary benefits to publication disproportionately benefit those researchers who are likely to get scooped. When startup costs are low, large sample-size scientists are more likely to get scooped. When startup costs are high, small sample-size scientists become increasingly likely to be scooped. However, large sample-size scientists have a higher probability of receiving the secondary payoff, conditional on being scooped, because they can achieve higher statistical power and have a higher probability of receiving the secondary payoff.

Discussion

Our initial findings support the hypothesis that competition for priority can reduce research quality: in our model, competition always causes scientists to reduce their sample size and statistical power. This occurs because scientists can increase their probability of being first by quickly conducting studies with small sample-sizes, even though this reduces the probability that any single study obtains a statistically-significant result. Larger numbers of competitors exacerbate this effect: when scientists compete against many other individuals, those scientists who strive to conduct studies with large sample-sizes often get scooped before they even get to run their study once. As a consequence, they are even more strongly incentivized to conduct small sample-size studies.

We also find that increased startup-costs allow populations to maintain higher sample sizes and statistical power at equilibrium. Startup costs are far from efficient: every researcher is forced to waste time in each investigation, resulting in populations of scientists that complete fewer studies (see Figure 5). However, startup costs de-incentivize a "quantity" strategy wherein researchers conduct large numbers of underpowered studies (Bakker et al., 2012) because these scientists experience startup costs more frequently.

Allowing scientists who get scooped to receive a non-zero payoff (i.e. rewarding secondary publications) also results in increased equilibrium sample sizes. In our model, this effect occurs whenever large sample-size scientists are most likely to be scooped. Allowing scooped scientists to receive some payoff reduces the incentive for scientists to

run small sample-size studies to increase their probability of being first. Our model thus provides theoretical support for the efficiency of "scoop protection" reforms at several academic journals (Editors, 2018; Marder, 2017; Yong, 2018a).

Although our model uses equilibrium sample size and statistical power as a proxy for research quality, it is not well-designed to determine what is harmful or beneficial for science as a whole. For example, we might be interested in whether increased competition leads scientists to publish research that has a higher ratio of false positives to true positives, as in other models of the scientific process (Ioannidis, 2005; Richard McElreath & Smaldino, 2015; Nissen et al., 2016; Smaldino & McElreath, 2016). The basic version of our model assumes that scientists only publish positive results, all scientists abandon a research question as soon as they are scooped by a competitor, and each research question is characterized by either no effect (i.e. $e = 0$) or a true effect (i.e. $e > 0$). As a consequence, all published results are positive, and the ratio of true positives to false positives is determined by the ratio of no-effect-to-true-effect research questions. To address this limitation, we are developing an extension that relaxes the aforementioned assumptions (e.g. both positive and negative results are published; scientists are not forced to abandon research questions upon being scooped; see https://github.com/ltiokhin/BESTEVERCompetitionModel/tree/Expansion_Abandonment_Bins_Pubbias). This will allow us to better determine the effect of competition and startup costs on science as a whole.

Our model has several assumptions that may be modified as we develop it further. We assume that the payoff for publication is independent of the sample size of a study or

effect size investigated. This assumption could be modified such that publishing large sample-size studies generates a higher payoff than publishing small sample-size studies, or that discovering a large effect results in a higher payoff than discovering a small effect. Several models of the scientific process have made different assumptions, including allowing the extent of publication bias to vary (Richard McElreath & Smaldino, 2015; Nissen et al., 2016) or allowing the probability of publication to depend on study quality (Higginson & Munafò, 2016). Our assumption of extreme publication bias (i.e. null results are not published, statistically-significant results are always published) corresponds better to some scientific fields than others. For instance, over 90% of published findings in Psychology/Psychiatry provide positive support for the hypothesis being tested, compared to approximately 70% in Space Science (Fanelli, 2010). Finally, our model assumes that researchers restart a project (i.e. throw away all their data) each time they test for statistical-significance but obtain a non-significant result, and that scientists cannot receive payoffs for intermediate results, unlike several other models of the scientific process (Bergstrom et al., 2016; T. Boyer, 2014).

Our finding that increased startup costs allow populations to maintain higher sample sizes and statistical power at equilibrium suggests that startup costs may be one viable solution to the problem of scientific reliability. Several proposals for scientific reform have already inadvertently introduced such startup costs. For example, pre-registration and registered reports make researchers spend more time thinking about and designing protocols before running investigations (Nosek, Ebersole, DeHaven, & Mellor,

2018; Nosek & Lakens, 2014). Although we currently conceptualize this time cost as an inconvenience, it may turn out to be key to incentivizing higher-quality research.

Open Practices Statement

All code for this project is available at:

https://github.com/ltiokhin/BESTEVERCompetitionModel

Ethics

Following the precedent set by (Smaldino & McElreath, 2016), all simulated scientists were humanely euthanized.

DISCUSSION

This combined research illustrates two aspects of scientific practice that may benefit from modification: the sample diversity of study participants and incentives for novel research findings. In the case of social discounting, an experiment conducted among the most diverse participant populations to date found that social discounting may not be a universal human phenomenon: there was no evidence that participants in rural Indonesia and rural Bangladesh were more generous to close social partners than distant ones. This finding could not be explained by various potential methodological issues, including floor effects, unreliable measurement of the independent or dependent variables, or lack of statistical power. Further, the fact that this experiment found the same pattern in Indonesia and Bangladesh, despite different protocols, provides convergent evidence that this finding is robust to methodological variations (Munafò & Smith, 2018). This provides evidence against the hypothesis that the null effect was a methodological artefact, and is consistent with the hypothesis that social discounting is a phenomenon that may only hold in a restricted range of human populations. It also illustrates the importance of complementing efforts at direct replication of published findings with investment in strong checks on generalizability across diverse samples. Without doing so, we risk developing theories of human nature that inevitably fail to generalize outside of the narrow range of participants on which most social-science research typically relies.

Although sample diversity has been largely overlooked by recent efforts at scientific reform (for example, see (Munafò et al., 2017)) incentive structures have

received more attention (Higginson & Munafò, 2016a; Munafò et al., 2017; Nosek et al., 2012). However, there is a dearth of empirical and theoretical evidence for precisely how different incentives affect the scientific process. Chapters 2 and 3 take one step towards addressing this gap by testing effects of incentivizing novel findings on information acquisition in an instantiation of the scientific process where the key decision that researchers face is how much data to collect before submitting an answer to a problem.

The experimental results in Chapter 2 demonstrate that the competition induced by rewarding novel findings can be harmful: individuals in the competition treatment made their guesses with less information (and as a result, had a higher probability of guessing incorrectly) than individuals in the no-competition treatment. The pilot study provided positive exploratory evidence for potential benefits of competition: individuals in the competition treatment solved arithmetic problems at a faster rate than individuals in the no-competition treatment. However, the subsequent high-powered confirmatory study provided no confirmatory evidence for the hypothesized benefits of competition: when individuals could adjust their effort to acquire more information, competition did not cause individuals to solve more arithmetic problems in order to acquire information more efficiently. This result is most likely if there was either 1) no effect of competition on effort in this particular study or 2) the effect was so small that the study was insufficiently powered to detect it.

The evolutionary agent-based model in Chapter 3 provides theoretical evidence that substantiates one of the main findings of the aforementioned experiment: competition for novel findings can incentivize individuals to rely on less evidence when

they guess the solution to research problems. In this model, competition for novel findings caused populations of scientists to evolve towards conducting research with smaller sample sizes and lower statistical power than when competition was absent. The model also provided theoretical evidence for the utility of two ways to increase research quality: startup costs and benefits for secondary publications. Higher time costs associated with setting up a single study allowed populations of scientists to maintain higher equilibrium sample sizes and statistical power. This occurred because high startup costs decreased the relative payoff to scientists who pursued a "quantity" strategy by conducting many low-quality studies. Increased benefits to secondary (i.e. non-novel results) also generally increased equilibrium sample size and statistical power. This occurred because individuals with large samples are often more likely to get scooped, and increasing benefits to secondary publications allows scooped individuals to receive at least some payoff for their research instead of receiving no benefit.

Limitations and Future Directions

The research in this dissertation is subject to a number of limitations. The cross-cultural study of social discounting (Chapter 1) is not able to provide a theoretical explanation for the finding of no social discounting among Bangladeshi and Indonesian participants. For instance, it is possible that social discounting only exists in a subset of human populations. Alternatively, it is possible that norms about whether individuals should behave according to personal preferences versus formal obligations determine whether or not the social distance between individuals affects their generosity (Miller & Bersoff, 1998). It is also possible that social discounting exists only given specific

experimental protocols (e.g. operationalizations of generosity), as slight differences in experimental protocol can sometimes generate dramatically different results (Landy et al., n.d.). Testing these and other explanations will be a fruitful direction for future research as scholars work towards a better understanding of the boundary conditions of social discounting.

The experimental test of how competition for novel results affects information sampling strategies (Chapter 2) is a useful first step to evaluate how incentive structures affect the reliability of science. However, it is subject to several limitations. The experiment is conducted with undergraduate participants, and it remains to be established whether the results will generalize to scientists. The experiment conceptualizes scientific discovery as effect hunting: there exist true effects which are independent of one another, and there is an identical payoff to discovering any given effect. These assumptions do not apply to some domains of scientific inquiry. Discoveries can vary in their payoff value, such that solving some research problems makes more of a scientific contribution than solving others, and scientists thereby receive a higher payoff for solving those problems (Bergstrom et al., 2016). Discoveries may also be interconnected, such that one finding increases the probability that scientists are able to make other findings, or make connections between previously unstudied phenomena (Rzhetsky, Foster, Foster, & Evans, 2015; Uzzi, Mukherjee, Stringer, & Jones, 2013). The fact that simple innovations allow individuals to discover more complex ones (Derex, Perreault, & Boyd, 2018) means that assuming independence between discoveries may not hold in real-world scientific practice.

Another limitation of the experiment is the fact that players in the Competition treatments compete solely against the performance of players in the No-Competition treatments. This results in a competitive situation where one individual's payoff (i.e. players in the Competition treatment) depends on their opponent's behavior, but an opponents' payoffs (i.e. players in the No-Competition treatment) do not depend on the behavior of other players. This causes a one-sided strategic interaction: players in Competition are expected to adjust their behavior based on their belief about the behavior of players in the No-Competition treatment, but players in the No-Competition treatment do not respond to the behavior of other players. As a result, the experiment provides a test of the direction of the effect of competition on information sampling (i.e. when people compete against individuals who are conducting research a non-competitive world, do people acquire more or less information before submitting their solution to the research problem?). The design does not provide information about equilibrium behavior (i.e. the stable long-run behavior when two individuals are in competition and can both adjust their behavior based on an expectation of the other player's likely behavior).

One-sided interaction could lead to different results than two-sided strategic interaction. In a one-sided game, players in the competition treatment are incentivized to guess slightly earlier than their opponent, while the opponent is not incentivized to change their behavior as a function of others play. In a two-sided game, both players are incentivized to guess earlier than each other, which could lead to a reduction in the number of tiles revealed across many rounds of play. In this case, a one-sided design would provide information about the direction of the effect of competition on tiles

revealed, while a two-sided design would be necessary to reveal equilibrium behavior. To illustrate the differences between the one-sided experimental design in Chapter 2 and that of other experiments on strategic interaction, I provide two examples of related experimental paradigms in behavioral game theory that are not one-sided (i.e. that allow both players to respond to each other's' behavior): the investment game and the "*p*-beauty contest" game.

In the symmetric version of the investment game, two players have an equal budget (*e*), and each have the option of making an irrecoverable investment to win an indivisible prize (*r*) (Camerer, 2011; Harris & Vickers, 1985). Players lose the proportion of their budget that is used to make their bid and they keep their remaining budget. The highest bidder receives the prize, but neither player receives the prize if both bid the same amount. This design conceptualizes bidding as an interaction where both players make a single, simultaneous decision, and each player's optimal decision depends on the decision of the other player. Consider the case where $e = 5$ and can potentially win $r = 8$. If their opponent invests 0, a player maximizes their payoff by investing $e = 1$, leaving them with a total payoff of 12. However, if an opponent invests everything (i.e. $e = 5$), a player maximizes their payoff by investing 0 (because investing everything would result in a bid equal to their opponent, resulting in 0 payoff), leaving them with a total payoff of 5. The bidding game has a unique symmetric mixed-strategy equilibrium, where individuals invest their whole endowment with probability $\frac{r-e}{r}$ and invest smaller integer amounts with equal probabilities $\frac{1}{r}$ (Camerer, 2011; Rapoport & Amaldoss, 2000). Experiments with undergraduate and graduate students provide evidence that the aggregate behavior of

individuals is close to the predicted mixed-strategy equilibrium (Camerer, 2011; Rapoport & Amaldoss, 2000).

The symmetric investment game has some similarities to the experiment in Chapter 2: interactions occur between 2 players and each player can make only a single decision per trial. One difference is that both players make their decisions simultaneously in the investment game, whereas decisions in the experiment in Chapter 2 have a temporal element: players get scooped if their competitor makes the correct guess faster than they do. This thus incentivizes players to anticipate how quickly their opponent is likely to make a guess and how accurate that guess will be. Another key difference is that the investment game incentivizes both players to anticipate the likely behavior of their competitor and adjust their own behavior accordingly, whereas the experiment in in Chapter 2 only allows one player to anticipate the other's behavior. Finally, the experiment in Chapter 2 allows players in the Competition treatment to learn about their competitor's behavior, whereas the investment game randomizes partner-pairings across after each trial, thereby only allowing players to learn about the average behavior of their many competitors (Rapoport & Amaldoss, 2000).

In the "p-beauty contest" game, each of n players simultaneously choose a number x in the interval 0 to 100. A multiple p of the average of all players' numbers is then chosen to be the "target number", and the player whose number x is closest to the target number wins a fixed prize (Camerer, 2011). The game was first described in (Moulin, 1986) as a way of measuring the number of steps of iterated reasoning engaged in by individuals. Assume that $p = 0.67$. In this game, people maximize their payoff by

guessing the average of other players' guesses, and then picking a number that is 2/3 of this average, knowing that all other players are also engaged in the same computation. This unique equilibrium strategy in this game is for players to guess 0. The reasoning is as follows. Players should never choose a number larger than 67: such a choice is dominated by choosing 67 (i.e. choosing 67 always results in a higher payoff than choosing a number greater than 67). If a player thinks that other individuals obey dominance, the player should choose *0.67 * 67 = 45*. But if all players think that everyone obeys one step of dominance and choose 45*,* then a player should choose *0.67 * 45 = 30*. Infinite steps of such iterated dominance lead to the unique equilibrium of 0.

Nagel conducted the first experimental test of player behavior in repeated "*p*-beauty contest" games (Nagel, 1995). She found that players typically exhibited 1-2 steps of iterated reasoning in first-round play. For example, when $p = 0.67$, most players guessed numbers between 20 and 40. In subsequent rounds of play, players' guesses decreases, such that guesses in the last round were closest to the predicted game-theoretic equilibrium of 0. Nagel's findings are corroborated by subsequent published studies of the "*p*-beauty contest" in western populations ranging from Cal-Tech undergraduates to high-school students in the U.S.: first-round choices are consistent with 1-3 steps of iterated reasoning and are far from the game-theoretic equilibrium, but players learn to make smaller guesses across multiple rounds of the experiment, such that the most guesses are 0 after 6-10 rounds of play (Camerer, 2011).

"*P*-beauty contest" games demonstrate that players' first round behavior may be far from predicted game-theoretic equilibria, even though players do move towards the

predicted equilibrium across multiple rounds of play. However, first-round play is not entirely misleading: players make guesses that are in the predicted direction (i.e. they choose numbers closer to 0 than would be expected if players choose at random). This is relevant to the experiment in Chapter 2, where only one player in a pair is incentivized to anticipate their competitor's behavior. That experimental design does not provide information about equilibrium behavior. However, it does provide information regarding the direction of the effect of competition on information sampling (i.e. when players A and B are competing, player B is rewarded for being first to accurately guess the underlying color, and player A is not anticipating the behavior of player B, does player B acquire more or less information?). An important future extension of the experiment in Chapter 2 would be to test equilibrium behavior by modifying the experiment such that both players can respond to competition for priority.

The evolutionary agent-based model in Chapter 3 has several assumptions that may limit its generality. It assumes that scientists are perfectly synchronized: all scientists start research questions at the same time, and as soon as one scientist publishes a positive result, all other scientists abandon their research question and move on to the subsequent one. This mirrors the experimental design of Chapter 2, but lacks realism. In more realistic competitive situations, finishing a problem early may allow scientists to get a head start on other research questions. The model also assumes that the payoff for publication is independent of the sample size of a study or effect size investigated. This assumption could be modified to make the payoff for publication contingent on research quality (e.g. scientists are rewarded more for publishing studies with large sample sizes).

Finally, the model assumes extreme publication bias: null results are never published, whereas statistically-significant results are always published. This approximates the current state of affairs in some fields better than others. For example, over 90% of published findings in Psychology/Psychiatry provide positive support for the hypothesis being tested, compared to approximately 70% in Space Science (Fanelli, 2010). An important future extension would be to test the effect of varying publication bias on the types of studies that scientists are incentivized to conduct. To address these limitations, a more general version of the model in Chapter 3 is currently being developed (see https://github.com/ltiokhin/BESTEVERCompetitionModel/tree/Expansion_Abandonment_Bins_Pubbias). This extension relaxes the assumption that scientists are always synchronized, allows both positive and negative results to be published, and does not force scientists to abandon a research question once a single result has been published on that question. This will allow an exploration of the extent to which the results of the model in Chapter 3 generalize beyond the specific assumptions of that model.

It is an exciting time to be a scientist. Scientists across disciplines are becoming aware that status quo scientific practice has problems that do not necessarily need to be accepted. We do not need to accept that individually-beneficial scientific practices necessarily conflict with the collective goals of scientists. We do not need to accept that publication requires a positive result that supports one's hypothesis, without showing any signs of uncertainty about the research (Frankenhuis & Nettle, 2018). And we do not need to accept a scientific world in which our theories are not formalized and our empirical work is conducted with populations that do not represent humanity as a whole.

The meta-scientific research in this dissertation takes one step towards addressing these issues. By itself, it may not have much impact. But in the aggregate, meta-scientific research provides our best hope for ensuring that scientists prioritize the pursuit of truth above all else.

REFERENCES

Alberts, B., Kirschner, M. W., Tilghman, S., & Varmus, H. (2014). Rescuing US biomedical research from its systemic flaws. *Proceedings of the National Academy of Sciences*, *111*, 5773–5777.

Anderson, M. S., Ronning, E. A., De Vries, R., & Martinson, B. C. (2007). The perverse effects of competition on scientists' work and relationships. *Science and Engineering Ethics*, *13*, 437–461.

Apicella, C. L., Azevedo, E. M., Christakis, N. A., & Fowler, J. H. (2014). Evolutionary origins of the endowment effect: Evidence from hunter-gatherers. *American Economic Review*, *104*, 1793–1805.

Apicella, C. L., & Barrett, H. C. (2016). Cross-cultural evolutionary psychology. *Current Opinion in Psychology*, *7*, 92–97.

Apicella, C. L., Crittenden, A. N., & Tobolsky, V. A. (2017). Hunter-gatherer males are more risk-seeking than females, even in late childhood. *Evolution and Human Behavior*, *38*, 592–603.

Arnett, J. J. (2008). The neglected 95%: why American psychology needs to become less American. *American Psychologist*, *63*, 602.

Aron, A., Aron, E. N., & Smollan, D. (1992). Inclusion of Other in the Self Scale and the structure of interpersonal closeness. *Journal of Personality and Social Psychology*, *63*, 596.

Aschw, C. (2015, August 19). Science Isn't Broken. Retrieved from https://fivethirtyeight.com/features/science-isnt-broken/

Baer, M., Vadera, A. K., Leenders, R. T., & Oldham, G. R. (2013). Intergroup competition as a double-edged sword: How sex composition regulates the effects of competition on group creativity. *Organization Science*, *25*, 892–908.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*, 452–454.

Baker, M. (n.d.). Dutch agency launches first grants programme dedicated to replication. *Nature News*. doi:10.1038/nature.2016.20287

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*, 543–554.

Balietti, S., Goldstone, R. L., & Helbing, D. (2016). Peer review and competition in the Art Exhibition Game. *Proceedings of the National Academy of Sciences*, 201603723.

Banerjee, S., Goel, A., & Kollagunta Krishnaswamy, A. (2014). Re-incentivizing discovery: Mechanisms for partial-progress sharing in research. In *Proceedings of the fifteenth ACM conference on Economics and computation* (pp. 149–166). ACM.

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., … Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences*, 201708285.

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M., Fitzpatrick, S., Gurven, M., … Pisor, A. (2016). Small-scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, *113*, 4688–4693.

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science. *Circulation Research*, *116*, 116–126.

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*, 407.

Benedictus, R., Miedema, F., & Ferguson, M. W. (2016). Fewer numbers, better science. *Nature*, *538*, 453–455.

Bergstrom, C. T., Foster, J. G., & Song, Y. (2016). Why scientists chase big problems: individual strategy and social optimality. *ArXiv Preprint ArXiv:1605.05822*. Retrieved from https://arxiv.org/abs/1605.05822

Bhattacharjee, Y. (2013, April 26). Diederik Stapel's Audacious Academic Fraud. *The New York Times*.

Boyer, P., Lienard, P., & Xu, J. (2012). Cultural differences in investing in others and in the future: why measuring trust is not enough. *PloS One*, *7*, e40750.

Boyer, T. (2014). Is a bird in the hand worth two in the bush? Or, whether scientists should publish intermediate results. *Synthese*, *191*, 17–35.

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., … van 't Veer, A. (2014). The Replication Recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224.

Brischoux, F., & Angelier, F. (2015). Academia's never-ending selection for productivity. *Scientometrics*, *103*, 333–336.

Bryant, G. A., Fessler, D. M. T., Fusaroli, R., Clint, E., Aarøe, L., Apicella, C. L., … Zhou, Y. (2016). Detecting affiliation in colaughter across 24 societies. *Proceedings of the National Academy of Sciences*, *113*, 4682–4687.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*, 365.

Camerer, C. F. (2011). *Behavioral game theory: Experiments in strategic interaction*. Princeton University Press.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., … others. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*, 1433–1436.

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. *Cortex*, *49*, 609–610.

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., … De Rosario, M. H. (2018). Package 'pwr.'

Clark, L., Robbins, T. W., Ersche, K. D., & Sahakian, B. J. (2006). Reflection impulsivity in current and former substance users. *Biological Psychiatry*, *60*, 515–522.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: a review. *The Journal of Abnormal and Social Psychology*, *65*, 145.

Coles, N., Tiokhin, L., Scheel, A. M., Isager, P. M., & Lakens, D. (2018). The Costs and Benefits of Replication Studies.

Collaboration, O. S., & others. (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716.

Cova, F., Strickland, B., Abatista, A. G. F., Allard, A., Andow, J., Attie, M., … Zhou, X. (2018). Estimating the Reproducibility of Experimental Philosophy. *PsyArXiv*. doi:10.17605/OSF.IO/SXDAH

Crandall, C. S., & Sherman, J. W. (2016). On the scientific superiority of conceptual replications for scientific progress. *Journal of Experimental Social Psychology*, *66*, 93–99.

Dechenaux, E., Kovenock, D., & Sheremeta, R. M. (2015). A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics*, *18*, 609–669.

Delton, A. W., & Robertson, T. E. (2016). How the mind makes welfare tradeoffs: Evolution, computation, and emotion. *Current Opinion in Psychology*, *7*, 12–16.

Derex, M., Perreault, C., & Boyd, R. (2018). Divide and conquer: intermediate levels of population fragmentation maximize cultural accumulation. *Phil. Trans. R. Soc. B*, *373*, 20170062.

Editors, T. P. B. S. (2018). The importance of being second. *PLOS Biology*, *16*, e2005203.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PloS One*, *5*, e10068.

Fanelli, D. (2011). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*, 891–904.

Fang, F. C., & Casadevall, A. (2015). Competitive Science: Is Competition Ruining Science? *Infection and Immunity*, *83*, 1229–1233.

Fiedler, K. (2011). Voodoo correlations are everywhere—not only in neuroscience. *Perspectives on Psychological Science*, *6*, 163–171.

Frankenhuis, W., & Nettle, D. (2018). Open Science is Liberating and Can Foster Creativity.

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances*, *3*, e1701381.

Geman, D., & Geman, S. (2016). Opinion: science in the age of selfies. *Proceedings of the National Academy of Sciences*, *113*, 9384–9387.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science*, *351*, 1037–1037.

Gneezy, U., Niederle, M., & Rustichini, A. (2003). Performance in competitive environments: Gender differences. *The Quarterly Journal of Economics*, *118*, 1049–1074.

Goeree, J. K., McConnell, M. A., Mitchell, T., Tromp, T., & Yariv, L. (2010). The 1/d law of giving. *American Economic Journal: Microeconomics*, *2*, 183–203.

Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*, 493–498.

Gurven, M., Stieglitz, J., Trumble, B., Blackwell, A. D., Beheim, B., Davis, H., … Kaplan, H. (2017). The tsimane health and life history project: Integrating anthropology and biomedicine. *Evolutionary Anthropology: Issues, News, and Reviews*, *26*, 54–73.

Hackman, J., Danvers, A., & Hruschka, D. J. (2015). Closeness is enough for friends, but not mates or kin: mate and kinship premiums in India and U.S. *Evolution and Human Behavior*, *36*, 137–145.

Hackman, J., Munira, S., Jasmin, K., & Hruschka, D. (2017). Revisiting Psychological Mechanisms in the Anthropology of Altruism. *Human Nature*, *28*, 76–91.

Harris, C., & Vickers, J. (1985). Patent races and the persistence of monopoly. *The Journal of Industrial Economics*, 461–481.

Heesen, R. (2017). Communism and the Incentive to Share in Science. *Philosophy of Science*, *84*, 698–716.

Henrich, J. (2015). Culture and social behavior. *Current Opinion in Behavioral Sciences*, *3*, 84–89.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., … Ensminger, J. (2005). "Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies. *Behavioral and Brain Sciences*, *28*, 795–815.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*, 61–83; discussion 83-135.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., … Henrich, N. (2006). Costly punishment across human societies. *Science*, *312*, 1767–1770.

Higginson, A. D., & Munafò, M. R. (2016). Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biology*, *14*, e2000995.

Hruschka, D. J., Munira, S., Jesmin, K., Hackman, J., & Tiokhin, L. (in press). Learning from failures of protocol in cross-cultural research. *Proceedings of the National Academy of Sciences*.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*, e124.

Ioannidis, J. P. (2014). How to make more published research true. *PLoS Medicine*, *11*, e1001747.

Ioannidis, J. P., Stanley, T. D., & Doucouliagos, H. (2017). The power of bias in economics research. *The Economic Journal*, *127*, F236–F265.

Ishii, K., & Eisen, C. (2018). Cultural Similarities and Differences in Social Discounting: The Mediating Role of Harmony-Seeking. *Frontiers in Psychology*, *9*. doi:10.3389/fpsyg.2018.01426

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*, 524–532.

Jones, B., & Rachlin, H. (2006). Social discounting. *Psychological Science*, *17*, 283–286.

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, *7*, 608–614.

Kruschke, J. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science*, 2515245918771304.

Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*, 355–362.

Lakens, D., Scheel, A. M., & Isager, P. M. (2017). Equivalence testing for psychological research: A tutorial.

Landy, J., et al. (n.d.). *Crowdsourcing hypothesis tests: Making transparent how design choices shape research results*.

Lezzi, E., Fleming, P., & Zizzo, D. J. (2015). Does it matter which effort task you use? a comparison of four effort tasks when agents compete for a prize.

Locey, M. L., Jones, B. A., & Rachlin, H. (2011). Real and hypothetical rewards. *Judgment and Decision Making*, *6*, 552.

Ma, Q., Pei, G., & Jin, J. (2015). What Makes You Generous? The Influence of Rural and Urban Rearing on Social Discounting in China. *PLOS ONE*, *10*, e0133078.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*, 537–542.

Marder, E. (2017). Scientific Publishing: Beyond scoops to best practices. *ELife*, *6*, e30076.

McElreath, R. (2012). Rethinking: statistical Rethinking book package. *R Package Version*, *1*.

McElreath, Richard, Bell, A. V., Efferson, C., Lubell, M., Richerson, P. J., & Waring, T. (2008). Beyond existence and aiming outside the laboratory: estimating frequency-dependent and pay-off-biased social learning strategies. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *363*, 3515–3528.

McElreath, Richard, & Smaldino, P. E. (2015). Replication, communication, and the population dynamics of scientific discovery. *PLoS One*, *10*, e0136088.

McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin*, *43*, 289–374.

Medin, D. L. (2017). Psychological science as a complex system: report card. *Perspectives on Psychological Science*, *12*, 669–674.

Merton, R. K. (1957). Priorities in scientific discovery: a chapter in the sociology of science. *American Sociological Review*, *22*, 635–659.

Miller, J. G., & Bersoff, D. M. (1998). The role of liking in perceptions of the moral responsibility to help: A cultural perspective. *Journal of Experimental Social Psychology*, *34*, 443–469.

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., … Protzko, J. (2018). Psychological Science Accelerator: Advancing Psychology through a Distributed Collaborative Network. *Advances in Methods and Practices in Psychological Science*.

Moulin, H. (1986). *Game theory for the social sciences*. NYU press.

Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., du Sert, N. P., … Ioannidis, J. P. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, 0021.

Munafò, M. R., & Smith, G. D. (2018). *Robust research needs many lines of evidence*. Nature Publishing Group.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *The American Economic Review*, *85*, 1313–1326.

Nelson, L. D., Simmons, J., & Simonsohn, U. (2017). Psychology's renaissance. *Annual Review of Psychology*. doi:https://doi.org/10.1146/annurev-psych-122216-011836

Nettle, D. (2017). *Tyneside neighbourhoods: Deprivation, social life and social behaviour in one British city*. Open Book Publishers.

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38.

Nissen, S. B., Magidson, T., Gross, K., & Bergstrom, C. T. (2016). Publication bias and the canonization of false facts. *Elife*, *5*, e21451.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 201708274.

Nosek, B. A., & Errington, T. M. (2017). Reproducibility in cancer biology: making sense of replications. *Elife*, *6*, e23383.

Nosek, B. A., & Lakens, D. (2014). Registered Reports. *Social Psychology*, *45*, 137–141.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*, 615–631.

Peterson, R. A. (2001). On the use of college students in social science research: Insights from a second-order meta-analysis. *Journal of Consumer Research*, *28*, 450–461.

Pornpattananangkul, N., Chowdhury, A., Feng, L., & Yu, R. (2017). Social discounting in the elderly: Senior citizens are good samaritans to strangers. *The Journals of Gerontology: Series B*.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, *10*, 712–712.

R Core Team. (2017). R: A language and environment for statistical computing. Retrieved from https://www.R-project.org/.

Rachlin, H., & Jones, B. A. (2008). Social discounting and delay discounting. *Journal of Behavioral Decision Making*, *21*, 29–43.

Rapoport, A., & Amaldoss, W. (2000). Mixed strategies and iterative elimination of strongly dominated strategies: An experimental investigation of states of knowledge. *Journal of Economic Behavior & Organization*, *42*, 483–521.

Rawat, S., & Meena, S. (2014). Publish or perish: Where are we heading? *Journal of Research in Medical Sciences : The Official Journal of Isfahan University of Medical Sciences*, *19*, 87–89.

Rozin, P. (2009). What kind of empirical research should we publish, fund, and reward?: A different perspective. *Perspectives on Psychological Science*, *4*, 435–439.

Rzhetsky, A., Foster, J. G., Foster, I. T., & Evans, J. A. (2015). Choosing experiments to accelerate collective discovery. *Proceedings of the National Academy of Sciences*, *112*, 14569–14574.

Sarewitz, D. (2016). The pressure to publish pushes down quality. *Nature*, *533*.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*, 461–464.

Schwarz, N., & Clore, G. L. (2016). Evaluating psychological research requires more than attention to the N: A comment on Simonsohn's (2015)"small telescopes." *Psychological Science*, *27*, 1407–1409.

Sears, D. O. (1986). College sophomores in the laboratory: Influences of a narrow data base on social psychology's view of human nature. *Journal of Personality and Social Psychology*, *51*, 515.

Silberzahn, R., Uhlmann, E. L., Martin, D., Anselmi, P., Aust, F., Awtrey, E. C., … Bonnier, E. (2017). Many analysts, one dataset: Making transparent how variations in analytical choices affect results.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*, 1359–1366.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017a). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123–1128.

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017b). Constraints on generality (COG): A proposed addition to all empirical papers. *Perspectives on Psychological Science*, *12*, 1123–1128.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*, 160384.

Strevens, M. (2003). The role of the priority rule in science. *The Journal of Philosophy*, *100*, 55–79.

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, *9*, 59–71.

Strombach, T., Jin, J., Weber, B., Kenning, P., Shen, Q., Ma, Q., & Kalenscher, T. (2014). Charity begins at home: Cultural differences in social discounting and generosity. *Journal of Behavioral Decision Making*, *27*, 235–245.

Strombach, T., Weber, B., Hangebrauk, Z., Kenning, P., Karipidis, I. I., Tobler, P. N., & Kalenscher, T. (2015). Social discounting involves modulation of neural value signals by temporoparietal junction. *Proceedings of the National Academy of Sciences*, *112*, 1619–1624.

The PLOS Biology Staff Editors. (2018). *The importance of being second*. Public Library of Science San Francisco, CA USA.

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*, 468–472.

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*, 6454–6459.

van Dijk, D., Manor, O., & Carey, L. B. (2014). Publication metrics and success on the academic job market. *Current Biology*, *24*, R516–R517.

Vekaria, K. M., Brethel-Haurwitz, K. M., Cardinale, E. M., Stoycos, S. A., & Marsh, A. A. (2017). Social discounting and distance perceptions in costly altruism. *Nature Human Behaviour*, *1*, s41562-017-0100–017.

Wagenmakers, E.-J. (2007). A Practical Solution to the Pervasive Problems of P Values. *Psychonomic Bulletin & Review*, *14*, 779.

Yong, E. (2018a). In Science, There Should Be a Prize for Second Place. Retrieved from https://www.theatlantic.com/science/archive/2018/02/in-science-there-should-be-a-prize-for-second-place/552131/

Yong, E. (2018b, February 1). In Science, There Should Be a Prize for Second Place. *The Atlantic*. Retrieved from https://www.theatlantic.com/science/archive/2018/02/in-science-there-should-be-a-prize-for-second-place/552131/

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*.

APPENDIX A

CHAPTER 1

PARTICIPANT DEMOGRAPHIC CHARACTERISTICS

	U.S.	**Bangladesh**	**Indonesia**
N	40	200	44
Age (mean(s.d.))	19.5 (1.3)	38.0 (14.3)	34.5 (9.6)
Sex (female/male (% female))	18/21 (46%)	166/200 (83%)	25/18 (58%)

Table 1S | Participant Demographic Characteristics. Two participants (one U.S., one Indonesian) reported their sex as "other" and were excluded from estimates of the male-female ratio.

MODEL COMPARISON

Testing the Best Model for All Sites

Model	DF	Log Likelihood	AIC	ΔAIC	BIC (Max N = 1388)	ΔBIC (Max N = 1388)	BIC (Min N = 284)	ΔBIC (Min N = 284)
Full Model (2 Ran. Slopes)	26	509.30	-966.59	0	-830.47	15.35	-871.73	8.99
No Ran. Slope Social Distance	22	502.5	-961	5.59	-845.82	0	-880.72	0
No Ran. Slope Need	19	395.97	-753.95	212.64	-654.47	191.35	-684.61	196.11
No Ran. Slope Social Distance or Need; Ran. Slope Relatedness	19	389.63	-741.27	225.32	-641.79	204.03	-671.93	208.79
Ran. Intercept Only	17	387.33	-740.65	225.94	-651.65	194.17	-678.63	202.09
No Ran. Effects	16	167.52	-303.05	663.54	-219.28	626.54	-244.66	636.06

Table 2S | Information criteria for different random-effect structures. Full Model (2 Ran. Slopes; Table 1, main text) includes fixed-effects for social distance, relatedness, and relative need, a random intercept for participant, and random slopes for both relative need and social distance. Table 2S compares this model to alternative models that differ only in their random-effects. No Ran. Slope Social Distance = random intercept for participant and random slope for need. No Ran. Slope Need = random intercept for participant and random slope for social distance. No Ran. Slope Social Distance or Need; Ran. Slope Relatedness = random intercept for participant and random slope for relatedness. Ran. Intercept only = random intercept for participant. No Ran. Effects = linear model with no random effects. 2 columns for Bayesian Information Criteria (BIC) indicate the upper and lower bounds on BIC. BIC with Max N = 1388 calculates BIC assuming each observation is independent. BIC with Min N = 284 calculates BIC assuming only 1 observation per participant (i.e. all observations for a given participant are entirely non-independent). ΔAIC and ΔBIC indicate the differences in information criteria between alternative models and the best model.

Model	DF	AIC	AIC Weight	BIC (Max N = 1388)	BIC Weight (Max N = 1388)	BIC (Min N = 284)	BIC Weight (Min N = 284)
Full Model (2 Ran. Slopes)	26	-966.59	0.94	-830.47	0	-871.73	0.01
Ran. Slope Need	22	-961	0.06	-845.82	1	-880.72	0.99
Ran. Slope Social Distance	19	-753.95	0	-654.47	0	-684.61	0
Ran. Slope Relatedness	19	-741.27	0	-641.79	0	-671.93	0
Ran. Intercept Only	17	-740.65	0	-651.65	0	-678.63	0
No Ran. Effects	16	-303.05	0	-219.28	0	-244.66	0

Table 3S | AIC and BIC Weights. Full Model (2 Ran. Slopes; Table 1, main text) includes fixed-effects for social distance, relatedness, and relative need, a random intercept for participant, and random slopes for both relative need and social distance. Table 3S compares the AIC and BIC weights of this model to alternative models that differ only in their random-effects (See Table 2S above)

Chi-Square Tests for Model Fit

Random intercept for participant is significant
Model 1: Fixed Effects (ln Social Distance, Relative Need, Relatedness); Random Effects (None)
Model 2: Fixed Effects (ln Social Distance, Relative Need, Relatedness); Random Effects (Random Intercept for Participant)

Model	**DF**	**ΔDF**	**Chisq**	**P-Value**
Model 1	16			
Model 2	17	1	439.61	<0.001

Random slope for social distance is significant
Model 2: Fixed Effects (ln Social Distance, Relative Need, Relatedness); Random Effects (Random Intercept for Participant)
Model 3: Fixed Effects (ln Social Distance, Relative Need, Relatedness); Random Effects (Random Intercept for Participant; Random Slope for ln Social Distance)

Model	**DF**	**ΔDF**	**Chisq**	**P-Value**
Model 2	17			
Model 3	19	2	17.3	<0.001

Random slope for relative need is significant
Model 2: Fixed Effects (ln Social Distance, Relative Need, Relatedness); Random Effects (Random Intercept for Participant)
Model 4: Fixed Effects (ln Social Distance, Relative Need, Relatedness); Random Effects (Random Intercept for Participant; Random Slope for Relative Need)

Model	**DF**	**ΔDF**	**Chisq**	**P-Value**
Model 2	17			
Model 4	22	5	230.35	<0.001

Random slope for relatedness is not significant
Model 2: Fixed Effects (ln Social Distance, Relative Need, Relatedness); Random Effects (Random Intercept for Participant)
Model 5: Fixed Effects (ln Social Distance, Relative Need, Relatedness); Random Effects (Random Intercept for Participant; Random Slope for Relatedness)

Model	**DF**	**ΔDF**	**Chisq**	**P-Value**
Model 2	17			
Model 5	19	2	4.62	0.10

A model with random slopes for both relative need and ln social distance is significantly better than a model with only a random slope for need or only a random slope for social distance
Model 3: Fixed Effects (ln Social Distance, Relative Need, Relatedness); Random Effects (Random Intercept for Participant; Random Slope for ln Social Distance)

Model 4: Fixed Effects (ln Social Distance, Relative Need, Relatedness); Random Effects (Random Intercept for Participant; Random Slope for Relative Need)
Model 6: Fixed Effects (ln Social Distance, Relative Need, Relatedness); Random Effects (Random Intercept for Participant; Random Slopes for ln Social Distance and Relative Need)

Model	**DF**	**ΔDF**	**Chisq**	**P-Value**
Model 3	19			
Model 6	26	7	226.65	<0.001

Model	**DF**	**ΔDF**	**Chisq**	**P-Value**
Model 4	22			
Model 6	26	4	13.60	0.009

ALTERNATIVE MODEL SPECIFICATIONS

All Sites

	U.S. Expected Sharing		Bangladesh Expected Sharing		Indonesia Expected Sharing	
	Estimate (CI)	*P*	*Estimate (CI)*	*P*	*Estimate (CI)*	*P*
Fixed Effects						
(Intercept)	0.70 (0.57 – 0.83)	<.001	0.15 (0.07 – 0.23)	<.001	0.66 (0.55 – 0.77)	<.001
Ln Social Distance	-0.10 (-0.12 – -0.08)	<.001	0.00 (-0.01 – 0.01)	.627	-0.00 (-0.02 – 0.01)	.620
Need						
Recipient Equally Needy	-0.10 (-0.22 – 0.03)	.140	-0.08 (-0.16 – 0.01)	.069	-0.19 (-0.29 – -0.09)	<.001
Recipient Less Needy	-0.19 (-0.32 – -0.06)	.004	-0.13 (-0.21 – -0.05)	.001	-0.30 (-0.41 – -0.19)	<.001
Relatedness	0.07 (-0.07 – 0.20)	.325	-0.01 (-0.10 – 0.08)	.866	0.13 (0.00 – 0.25)	.049

Table 4S| Generosity as a function of social distance, relative need, and relatedness. Only random slope for relative need. Multilevel model of social distance, recipient need, and relatedness regressed on expected sharing. Model controls for correlated observations from the same participant with random effects for each individual and includes a random slope for recipient need. CI = 95% confidence intervals.

	U.S. Expected Sharing		Bangladesh Expected Sharing		Indonesia Expected Sharing	
	Estimate (CI)	*P*	*Estimate (CI)*	*P*	*Estimate (CI)*	*P*
Fixed Effects						
(Intercept)	0.71 (0.58 – 0.84)	<.001	0.15 (0.07 – 0.22)	<.001	0.70 (0.59 – 0.80)	<.001
Ln Social Distance	-0.10 (-0.12 – -0.08)	<.001	0.00 (-0.01 – 0.01)	.661	-0.01 (-0.03 – 0.01)	.226
Need						
Recipient Equally Needy	-0.10 (-0.22 – 0.02)	.117	-0.07 (-0.15 – 0.01)	.087	-0.20 (-0.30 – -0.10)	<.001
Recipient Less Needy	-0.19 (-0.32 – -0.07)	.004	-0.13 (-0.21 – -0.05)	.001	-0.31 (-0.42 – -0.20)	<.001

Table 5S| Generosity as a function of social distance and relative need (excluding genetic relatedness). Multilevel model of social distance and recipient need regressed on expected sharing. Model controls for correlated observations from the same participant with random effects for each individual and includes random slopes for social distance and recipient need. CI = 95% confidence intervals.

	U.S. Expected Sharing		Bangladesh Expected Sharing		Indonesia Expected Sharing	
	Estimate (CI)	*P*	*Estimate (CI)*	*P*	*Estimate (CI)*	*P*
Fixed Effects						
(Intercept)	0.61 (0.54 – 0.67)	<.001	0.04 (0.02 – 0.07)	.003	0.52 (0.45 – 0.59)	<.001
Ln Social Distance	-0.11 (-0.13 – -0.09)	<.001	0.00 (-0.01 – 0.01)	.875	-0.03 (-0.05 – -0.01)	.013
Relatedness	0.04 (-0.12 – 0.19)	.653	-0.04 (-0.15 – 0.07)	.453	0.14 (0.00 – 0.27)	.046

Table 6S| Generosity as a function of social distance and relatedness (excluding need). Multilevel model of social distance and relatedness regressed on expected sharing. Model controls for correlated observations from the same participant with random effects for each individual and includes a random slope for social distance. Without controlling for need, social distance has a stronger estimated association with generosity in Indonesia. CI = 95% confidence intervals.

	U.S. Expected Sharing		Bangladesh Expected Sharing		Indonesia Expected Sharing	
	Estimate (CI)	*P*	*Estimate (CI)*	*P*	*Estimate (CI)*	*P*
Fixed Effects						
(Intercept)	0.61 (0.55 – 0.68)	<.001	0.04 (0.01 – 0.07)	.004	0.55 (0.49 – 0.61)	<.001
Ln Social Distance	-0.11 (-0.14 – -0.09)	<.001	0.00 (-0.01 – 0.01)	.775	-0.04 (-0.06 – -0.02)	<.001

Table 7S| Generosity as a function of social distance (excluding need and genetic relatedness). Multilevel model of social distance regressed on expected sharing. Model controls for correlated observations from the same participant with random effects for each individual and includes a random slope for social distance. When removing all co-variates, social distance has a stronger estimated association with generosity in Indonesia. CI = 95% confidence intervals.

	Expected Sharing	
	Estimate (CI)	*P*
Fixed Effects		
(Intercept)	0.71 (0.57 – 0.84)	<.001
Ln Social Distance (U.S.)	-0.10 (-0.12 – -0.08)	<.001
Site (Reference = U.S)		
Bangladesh	-0.56 (-0.71 – -0.41)	<.001
Indonesia	-0.04 (-0.21 – 0.13)	.670
Relative Need		
Recipient Equally Needy	-0.10 (-0.22 – 0.02)	.121
Recipient Less Needy	-0.19 (-0.32 – -0.07)	.004
Relatedness	0.05 (-0.08 – 0.19)	.459
Ln Social Distance: Bangladesh	0.10 (0.08 – 0.12)	<.001
Ln Social Distance: Indonesia	0.10 (0.07 – 0.12)	<.001
Bangladesh: Recipient Equally Needy	0.03 (-0.12 – 0.17)	.721
Indonesia: Recipient Equally Needy	-0.10 (-0.25 – 0.06)	.224
Bangladesh: Recipient Less Needy	0.06 (-0.09 – 0.21)	.401

Indonesia: Recipient Less Needy	-0.11 (-0.28 – 0.06)	.191
Bangladesh: Relatedness	-0.06 (-0.22 – 0.10)	.475
Indonesia: Relatedness	0.07 (-0.11 – 0.25)	.452
Random Parts		
σ^2	0.016	
$\tau_{00,\ respid}$	0.080	
ρ_{01}	-0.750	
N_{respid}	284	
Observations	1388	
R^2 / Ω_0^2	.882 / .879	

Table 8S| Generosity as a function of social distance, relative need, and relatedness (Full output for Table 1 in main text). Multilevel model of social distance, relative need, and relatedness regressed on expected sharing. Model controls for correlated observations from the same participant with random effects for each individual and includes random slopes for social distance and relative need. Model compares effect estimates in Bangladesh and Indonesia to the U.S. (i.e. the reference group). CI = 95% confidence intervals.

	U.S. Expected Sharing		Bangladesh Expected Sharing		Indonesia Expected Sharing	
	Estimate (CI)	*P*	*Estimate (CI)*	*P*	*Estimate (CI)*	*P*
Fixed Effects						
(Intercept)	0.68 (0.55 – 0.80)	<.001	0.15 (0.07 – 0.23)	<.001	0.67 (0.57 – 0.78)	<.001
Social Distance	-0.02 (-0.02 – -0.01)	<.001	0.00 (-0.00 – 0.00)	.618	-0.00 (-0.01 – 0.00)	.315
Need						
Recipient Equally Needy	-0.11 (-0.23 – 0.02)	.095	-0.08 (-0.16 – 0.00)	.064	-0.20 (-0.30 – -0.10)	<.001
Recipient Less Needy	-0.20 (-0.33 – -0.08)	.002	-0.13 (-0.21 – -0.06)	<.001	-0.30 (-0.41 – -0.19)	<.001
Relatedness	0.04 (-0.09 – 0.18)	.548	-0.01 (-0.10 – 0.08)	.870	0.11 (-0.01 – 0.23)	.081

Table 9S| Generosity as a function of social distance (unlogged), relative need, and relatedness. Multilevel model of raw (unlogged) social distance, relative need, and relatedness regressed on expected sharing. Model controls for correlated observations from the same participant with random effects for each individual and includes random slopes for social distance and relative need. CI = 95% confidence intervals.

Within Sites

	Bangladesh Expected Sharing	
	Estimate (CI)	*P*
Fixed Effects		
(Intercept)	0.13 (0.05 – 0.22)	.002
Ln Social Distance	0.00 (-0.00 – 0.01)	.444
Relative Need		
Recipient Equally Needy	-0.07 (-0.15 – 0.01)	.106
Recipient Less Needy	-0.12 (-0.20 – -0.04)	.004
Relatedness	-0.01 (-0.06 – 0.04)	.799
Order_Asked		
Order_Asked2	0.01 (-0.00 – 0.03)	.120
Order_Asked3	-0.01 (-0.02 – 0.01)	.260
Order_Asked4	-0.00 (-0.02 – 0.01)	.718
Order_Asked5	0.00 (-0.02 – 0.02)	.954
Age	0.00 (-0.00 – 0.00)	.914
Recipient Gender (Reference Category = Female)		

Male	0.00 (-0.01 – 0.01)	.911
Unspecified	0.01 (-0.16 – 0.17)	.949
Participant Gender (Reference Category = Female)	0.01 (-0.01 – 0.04)	.343
Random Parts		
σ^2	0.005	
$\tau_{00,\ respid}$	0.088	
ρ_{01}	0.077	
N_{respid}	200	
Observations	964	
R^2 / Ω_0^2	.876 / .874	

Table 10S| Generosity among Bangladesh participants as a function of social distance, relative need, and relatedness, controlling for participant and recipient gender, order of recipient, and participant age. Multilevel model of social distance, recipient need, and relatedness regressed on expected sharing. Model also includes fixed effects for participant gender, recipient gender, order or recipient, and participant age. Model controls for correlated observations from the same participant with random effects for each individual and includes random slopes for social distance and recipient need. CI = 95% confidence intervals.

	Indonesia Expected Sharing	
	Estimate (CI)	*P*
Fixed Effects		
(Intercept)	0.69 (0.46 – 0.93)	<.001
Ln Social Distance	-0.00 (-0.03 – 0.03)	.871
Relative Need		
Recipient Equally Needy	-0.23 (-0.33 – -0.13)	<.001
Recipient Less Needy	-0.34 (-0.46 – -0.22)	<.001
Relatedness	0.12 (-0.07 – 0.31)	.210
Order_Asked		
Order_Asked2	0.03 (-0.06 – 0.12)	.464
Order_Asked3	-0.01 (-0.10 – 0.08)	.780
Order_Asked4	0.03 (-0.07 – 0.12)	.589
Order_Asked5	-0.02 (-0.12 – 0.07)	.623
Order_Asked6	0.08 (-0.19 – 0.36)	.569
Age	0.00 (-0.01 – 0.01)	.783

Recipient Gender (Reference Category = Female)		
Male	-0.03 (-0.11 – 0.04)	.402
Unspecified	0.06 (-0.10 – 0.21)	.464
Participant Gender (Reference Category = Female)	-0.05 (-0.17 – 0.07)	.436
Random Parts		
σ^2	0.039	
$\tau_{00,\ respid}$	0.027	
ρ_{01}	-0.273	
N_{respid}	43	
Observations	215	
R^2 / Ω_0^2	.737 / .722	

Table 11S| Generosity among Indonesian participants as a function of social distance, relative need, and relatedness, controlling for participant and recipient gender, order of recipient, and participant age. Multilevel model of social distance, recipient need, and relatedness regressed on expected sharing. Model also includes fixed effects for participant gender, recipient gender, order of recipient, and participant age. Model controls for correlated observations from the same participant with random effects for each individual and includes a random slope for recipient need. CI = 95% confidence intervals.

	U.S. Expected Sharing	
	Estimate (CI)	*P*
Fixed Effects		
(Intercept)	0.58 (-0.68 – 1.84)	.376
Ln Social Distance	-0.11 (-0.14 – -0.08)	<.001
Relative Need		
Recipient Equally Needy	-0.13 (-0.24 – -0.01)	.032
Recipient Less Needy	-0.22 (-0.33 – -0.11)	<.001
Relatedness	0.14 (-0.09 – 0.37)	.223
Order_Asked		
Order_Asked2	-0.13 (-0.23 – -0.02)	.020
Order_Asked3	-0.08 (-0.19 – 0.03)	.145
Order_Asked4	-0.20 (-0.31 – -0.10)	<.001
Order_Asked5	-0.11 (-0.21 – -0.00)	.051
Order_Asked6	0.01 (-0.10 – 0.12)	.808
Age	0.01 (-0.05 – 0.08)	.729

Recipient Male (Reference Category = Female)	0.04 (-0.03 – 0.12)	.284
Participant Male (Reference Category = Female)	0.02 (-0.15 – 0.19)	.842
Random Parts		
σ^2	0.047	
$\tau_{00,\ respid}$	0.061	
N_{respid}	39	
Observations	195	
R^2 / Ω_0^2	.728 / .722	

Table 12S| Generosity among U.S. participants as a function of social distance, relative need, and relatedness, controlling for participant and recipient gender, order, and participant age. Multilevel model of social distance, recipient need, and relatedness regressed on expected sharing. Model also includes fixed effects for participant gender, recipient gender, order of recipient, and participant age. Model controls for correlated observations from the same participant with random effects for each individual. CI = 95% confidence intervals.

BIC AND BAYES FACTORS FOR MODELS WITH/WITHOUT SOCIAL DISTANCE

We calculate Bayesian Information Criterion (BIC)(G. Schwarz, 1978) values to assess the extent to which the data favor models (i.e. statistical descriptions of hypotheses) with or without social distance. We then use BIC values to approximate Bayes Factors (BF) for competing models(Wagenmakers, 2007).

Bangladesh	**DF**	**Log Likelihood**	**BIC (Max N = 968)**	**BIC (Min N = 200)**
Full Model (2 Ran. Slopes)	16	902.25	-1694.5	-1719.73
No Ran. Slope Social Distance	12	881.47	-1680.4	-1699.36
No Ran. Slope or Fixed Effect Social Distance	11	881.06	-1686.5	-1703.84
Indonesia	**DF**	**Log Likelihood**	**BIC (Max N = 220)**	**BIC (Min N = 44)**
Full Model (2 Ran. Slopes)	16	-6.31	98.92	73.16
No Ran. Slope Social Distance	12	-8.02	80.77	61.45
No Ran. Slope or Fixed Effect Social Distance	11	-8.08	75.49	57.78
U.S.	**DF**	**Log Likelihood**	**BIC (Max N = 200)**	**BIC (Min N = 40)**
Full Model (2 Ran. Slopes)	16	-25.34	135.45	109.7
No Ran. Slope Social Distance	12	-26.74	117.05	97.75
No Ran. Slope or Fixed Effect Social Distance	11	-46.28	150.85	133.14

Table 13S | BIC for competing models. Full Model (2 Ran. Slopes; Table 1, main text) includes fixed-effects for social distance, relatedness, and relative need, a random intercept for participant, and random slopes for both relative need and social distance. This table compares this model to alternative models that differ by removing just the random slope for social distance (No Ran. Slope Social Distance) or by removing both the random and fixed effect of social distance (No Ran. Slope or Fixed Effect Social Distance). 2 columns for Bayesian Information Criteria (BIC) indicate the upper and lower bounds on BIC. BIC with Max N calculates BIC assuming each observation is independent. BIC with Min N calculates BIC assuming only 1 observation per participant (i.e. all observations for a given participant are entirely non-independent).

We approximate Bayes Factors (BF) by exponentiating half the difference between the BIC values of competing models (i.e. $\exp(\Delta BIC_{10} / 2)$).(Wagenmakers, 2007) BF_{10} indicates a ratio: the likelihood of the data conditional on Model 1, $P(D|M_1)$, divided by the likelihood of the data conditional on Model 0, $P(D|M_0)$. For example, if $BF_{10} = 8$, the data are 8 times more likely under Model 1 than Model 0. If $BF_{10} = 0.01$, the

data are 100 times less likely under Model 1 than Model 0. For all below comparisons, Model 1 is listed first and Model 0 is listed second.

Bangladesh BF
Full Model (2 Ran. Slopes) vs. No Ran. Slope Social Distance
Using BIC Max N. $BF_{10} = 1152.86$
Using BIC Min N. $BF_{10} = 26462.93$
Full Model (2 Ran. Slopes) vs. No Ran. Slope or Fixed Effect Social Distance
Using BIC Max N. $BF_{10} = 54.60$
Using BIC Min N. $BF_{10} = 2818.61$
No Ran. Slope Social Distance vs. No Ran. Slope or Fixed Effect Social Distance
Using BIC Max N. $BF_{10} = 0.05$
Using BIC Min N. $BF_{10} = 0.11$

Indonesia BF
Full Model (2 Ran. Slopes) vs. No Ran. Slope Social Distance
Using BIC Max N. $BF_{10} = 0.0001$
Using BIC Min N. $BF_{10} = 0.0029$
Full Model (2 Ran. Slopes) vs. No Ran. Slope or Fixed Effect Social Distance
Using BIC Max N. $BF_{10} = 0.000008$
Using BIC Min N. $BF_{10} = 0.00046$
No Ran. Slope Social Distance vs. No Ran. Slope or Fixed Effect Social Distance
Using BIC Max N. $BF_{10} = 0.07$
Using BIC Min N. $BF_{10} = 0.16$

U.S. BF
Full Model (2 Ran. Slopes) vs. No Ran. Slope Social Distance
Using BIC Max N. $BF_{10} = 0.0001$
Using BIC Min N. $BF_{10} = 0.0026$
Full Model (2 Ran. Slopes) vs. No Ran. Slope or Fixed Effect Social Distance
Using BIC Max N. $BF_{10} = 2208.35$
Using BIC Min N. $BF_{10} = 123007.4$
No Ran. Slope Social Distance vs. No Ran. Slope or Fixed Effect Social Distance
Using BIC Max N. $BF_{10} = 21856305$
Using BIC Min N. $BF_{10} = 48399498$

In Bangladesh and Indonesia, BF indicate support for a model without a fixed effect for social distance, whereas in the U.S., BF indicate support for a model with a fixed effect for social distance. In Indonesia and the U.S., BF also indicate support for a model without a random slope for social distance, whereas in Bangladesh, BF indicate support for a model with a random slope for social distance.

EXCLUSIONS AND INCLUSIONS

Excluding Participants Who Gave Nothing to All Recipients

	U.S. Expected Sharing		Bangladesh Expected Sharing		Indonesia Expected Sharing	
	Estimate (CI)	*P*	*Estimate (CI)*	*P*	*Estimate (CI)*	*P*
Fixed Effects						
(Intercept)	0.74 (0.59 – 0.89)	<.001	0.41 (0.28 – 0.55)	<.001	0.69 (0.56 – 0.82)	<.001
Ln Social Distance	-0.10 (-0.14 – -0.07)	<.001	0.01 (-0.03 – 0.04)	.615	-0.01 (-0.04 – 0.03)	.715
Need						
Recipient Equally Needy	-0.12 (-0.26 – 0.03)	.118	-0.15 (-0.30 – -0.00)	.052	-0.20 (-0.32 – -0.08)	.001
Recipient Less Needy	-0.21 (-0.36 – -0.06)	.007	-0.30 (-0.43 – -0.16)	<.001	-0.31 (-0.44 – -0.18)	<.001
Relatedness	0.05 (-0.16 – 0.26)	.630	-0.03 (-0.36 – 0.31)	.878	0.13 (-0.07 – 0.32)	.202
Random Parts						
σ^2	0.042					
$\tau_{00,\ respid}$	0.077					
ρ_{01}	-0.692					
N_{respid}	116					
Observations	576					
R^2 / Ω_0^2	.781 / .768				.	

Table 14S| Generosity as a function of social distance, relative need, and relatedness, only including participants with non-zero generosity. Multilevel model of social distance, recipient need, and relatedness regressed on expected sharing. Model controls for correlated observations from the same participant

with random effects for each individual and includes random slopes for social distance and recipient need. When excluding participants who gave nothing to all recipients, the effect of social distance on generosity remains largely unchanged in each site. Number of participants = 35 (Bangladesh), 39 (U.S.), 42 (Indonesia). CI = 95% confidence intervals.

Including Participant Decisions Towards "Unknown Person"

To assess individual decisions unfamiliar partners, participants in all sites also made decisions between selfish and generous options for an "unknown person". Below, we reanalyze the data, including generosity towards an "unknown person".

	U.S. Expected Sharing		Bangladesh Expected Sharing		Indonesia Expected Sharing	
	Estimate (CI)	*P*	*Estimate (CI)*	*P*	*Estimate (CI)*	*P*
Fixed Effects						
Intercept for "Unknown Individual"	0.39 (0.28, 0.50)	<.001	0.08 (0.03, 0.13)	.003	0.60 (0.51, 0.70)	<.001
Change in Intercept for Recipients with a Social Distance	0.36 (0.30, 0.42)	<.001	0.01 (-0.02, 0.04)	.381	0.03 (-0.03, 0.09)	.291
Ln Social Distance	-0.10 (-0.12, -0.08)	<.001	0.00 (-0.01, 0.01)	.605	-0.00 (-0.02, 0.02)	.943
Need						
Recipient Equally Needy	-0.15 (-0.26, -0.05)	.003	-0.03 (-0.08, 0.02)	.257	-0.16 (-0.25, -0.08)	<.001
Recipient Less Needy	-0.24 (-0.34, -0.13)	<.001	-0.07 (-0.13, -0.02)	.006	-0.30 (-0.39, -0.21)	<.001

Relatedness	0.06 (-0.07, 0.19)	.335	-0.02 (-0.10, 0.07)	.722	0.16 (0.04, 0.28)	.010
Random Parts						
σ^2	0.018					
$\tau_{00,\ respid}$	0.057					
ρ_{01}	0.086					
N_{respid}	284					
Observations	1671					
R^2 / Ω_0^2	.858 / .855					

Table 15S| Generosity as a function of social distance, relative need, and relatedness, including data for generosity towards an "unknown person". Multilevel model of social distance, recipient need, relatedness, and whether recipient had a social-distance ranking (categorical) regressed on expected sharing. Model controls for correlated observations from the same participant with random effects for each individual and includes random slopes for social distance and recipient need. 'Intercept for "Unknown Individual"' indicates the estimate for participant generosity towards an "unknown individual". 'Change in Intercept for Recipients with a Social Distance" indicates the change in intercept, relative to 'Intercept for "Unknown Individual"', for recipients who have a social-distance ranking (i.e. the expected sharing towards a recipient at social distance = 1). Number of participants = 200 (Bangladesh), 40 (U.S.), 44 (Indonesia). CI = 95% confidence intervals.

GENEROSITY AS A FUNCTION OF PAYOFF TO PARTICIPANT

	U.S. Odds of Sharing		Bangladesh Odds of Sharing		Indonesia Odds of Sharing	
	Estimate (CI)	*P*	*Estimate (CI)*	*P*	*Estimate (CI)*	*P*
Fixed Effects						
(Intercept)	0.17 (0.08, 0.38)	<.001	0.00 (0.00, 0.00)	<.001	0.47 (0.29, 0.78)	.003
Decision						
2	2.62 (1.51, 4.55)	<.001	1.56 (0.81, 3.02)	.184	2.27 (1.47, 3.50)	<.001
3	6.76 (3.86, 11.84)	<.001	1.73 (0.90, 3.32)	.101	3.33 (2.15, 5.16)	<.001
4	16.88 (9.39, 30.36)	<.001	3.65 (1.94, 6.86)	<.001	2.88 (1.86, 4.45)	<.001
5	39.17 (20.81, 73.71)	<.001	7.10 (3.80, 13.26)	<.001	2.74 (1.78, 4.24)	<.001
6	88.42 (43.73, 178.78)	<.001	12.61 (6.72, 23.67)	<.001	3.50 (2.25, 5.43)	<.001
Random Parts						
$\tau_{00,\ respid}$	4.503		85.830		1.713	
N_{respid}	40		200		44	
Observations	1200		5808		1320	
Deviance	908.292		876.467		1417.457	

Table 16S| Generosity as a function of payoff to participant. Logistic regression of payoff to participant (i.e. cost of sharing) regressed on sharing (binomial yes/no outcome). Larger numbers for "Decision" indicate smaller participant payoffs (i.e. smaller costs to sharing). Model controls for correlated observations from the same participant with random effects for each individual. Excludes data for generosity towards "unknown person" and "acquaintance". In all sites, participants have higher odds of sharing on decisions when the personal costs of doing so are low.

INTERACTION BETWEEN PARTICIPANT CONSISTENSY AND GENEROSITY

	U.S. Expected Sharing	
	Estimate (CI)	*P*
Fixed Effects		
(Intercept)	0.74 (0.52, 0.96)	<.001
Ln Social Distance	-0.10 (-0.16, -0.03)	.004
Consistent Participant	0.03 (-0.19, 0.25)	.778
Need		
Recipient Equally Needy	-0.13 (-0.25, -0.01)	.036
Recipient Less Needy	-0.22 (-0.33, -0.11)	<.001
Relatedness	0.08 (-0.15, 0.31)	.482
Ln Social Distance * Consistent Participant Interaction	-0.01 (-0.09, 0.06)	.687
Random Parts		
σ^2	0.055	
$\tau_{00,\ respid}$	0.055	
N_{respid}	39	
Observations	195	
R^2 / Ω_0^2	.677 / .669	

Table 17S| Generosity among U.S. participants as a function of social distance, relative need, relatedness, and participant consistency, only including participants with non-zero generosity. Multilevel model of social distance, recipient need, relatedness, and participant consistency (categorical: 1, 0) regressed on expected sharing. Participants were considered inconsistent if they had multiple crossover points for at least 1 recipient. Model controls for correlated observations from the same participant with random effects for each individual and includes random slopes for social distance and recipient need. CI = 95% confidence intervals.

	Bangladesh Expected Sharing	
	Estimate (CI)	*P*
Fixed Effects		
(Intercept)	0.40 (0.24, 0.56)	<.001
Ln Social Distance	0.01 (-0.02, 0.05)	.420
Consistent Participant	0.07 (-0.10, 0.25)	.424
Need		
Recipient Equally Needy	-0.16 (-0.34, 0.03)	.106
Recipient Less Needy	-0.30 (-0.47, -0.14)	.002
Relatedness	-0.03 (-0.34, 0.28)	.847
Ln Social Distance * Consistent Participant Interaction	-0.01 (-0.08, 0.06)	.760
Random Parts		
σ^2	0.035	
$\tau_{00,\ respid}$	0.108	
ρ_{01}	-0.767	
N_{respid}	35	
Observations	171	
R^2 / Ω_0^2	.782 / .774	

Table 18S| Generosity among Bangladesh participants as a function of social distance, relative need, relatedness, and participant consistency, only including participants with non-zero generosity. Multilevel model of social distance, recipient need, relatedness, and participant consistency (categorical: 1, 0) regressed on expected sharing. Participants were considered inconsistent if they had multiple crossover points for at least 1 recipient. Model controls for correlated observations from the same participant with random effects for each individual and includes a random slope for recipient need. CI = 95% confidence intervals.

BIC AND BAYES FACTORS FOR MODELS WITH/WITHOUT PARTICIPANT-CONSISTENCY INTERACTION

We calculate Bayesian Information Criterion (BIC)(G. Schwarz, 1978) values to assess the extent to which the data favor models (i.e. statistical descriptions of hypotheses) with or without an interaction between Ln Social Distance and Participant Consistency in Bangladesh and the U.S.. We then use BIC values to approximate Bayes Factors (BF) for competing models.

U.S.	**DF**	**Log Likelihood**	**BIC (Max N = 195)**	**BIC (Min N = 39)**
Interaction Model (Ran. Intercept)	9	-28.12	98.59	85.73
No-Interaction Model (Ran. Intercept)	8	-28.21	103.71	89.21
Bangladesh	**DF**	**Log Likelihood**	**BIC (Max N = 171)**	**BIC (Min N = 35)**
Interaction Model (Ran. Intercept + 1 Ran. Slope)	14	1.52	68.86	46.65
No-Interaction Model (Ran. Intercept + 1 Ran. Slope)	13	1.56	63.81	43.18

Table 19S | BIC for competing models. In the U.S., Interaction Model (Ran. Intercept) includes fixed-effects for social distance, relatedness, relative need, and participant consistency (categorical; 1, 0), an interaction between participant consistency and social distance, and a random intercept for participant. In Bangladesh, Interaction Model (Ran. Intercept + 1 Ran. Slope) is an identical model, but also includes a random slope for recipient need. Both models only include data from participants with non-zero generosity. In both sites, the No-Interaction model removes the interaction between participant consistency and social distance. Participants were considered inconsistent if they had multiple crossover points for at least 1 recipient. 2 columns for Bayesian Information Criteria (BIC) indicate the upper and lower bounds on BIC. BIC with Max N

calculates BIC assuming each observation is independent. BIC with Min N calculates BIC assuming only 1 observation per participant (i.e. all observations for a given participant are entirely non-independent).

We approximate Bayes Factors (BF) by exponentiating half the difference between the BIC values of competing models (i.e. $\exp(\Delta BIC_{10} / 2)$).(Wagenmakers, 2007) BF_{10} indicates a ratio: the likelihood of the data conditional on Model 1, $P(D|M_1)$, divided by the likelihood of the data conditional on Model 0, $P(D|M_0)$. For example, if $BF_{10} = 8$, the data are 8 times more likely under Model 1 than Model 0. If $BF_{10} = 0.01$, the data are 100 times less likely under Model 1 than Model 0. For all below comparisons, Model 1 is listed first and Model 0 is listed second.

U.S. BF
No-Interaction Model (Ran. Intercept) vs Interaction Model (Ran. Intercept)
Using BIC Max N. $BF_{10} = 12.94$
Using BIC Min N. $BF_{10} = 5.70$

Bangladesh BF
No-Interaction Model (Ran. Intercept + 1 Ran. Slope) vs Interaction Model (Ran. Intercept + 1 Ran. Slope)
Using BIC Max N. $BF_{10} = 12.49$
Using BIC Min N. $BF_{10} = 5.67$

In both Bangladesh and the U.S., BF indicate support for a model without an interaction between social distance and participant consistency.

POWER ANALYSIS

Power of the current study to detect a significant independent effect of Social Distance on Expected Sharing, as a function of effect size. Analysis uses the powerSim() function from the SIMR(Green & MacLeod, 2016) package in R(R Core Team, 2017). For each simulation, powerSim() simulates new values for Expected Sharing from the model in Table 1 (main text), refits this model using those values, and performs a two-sized z-test on the simulated data. Analysis based on 1000 simulations ($\alpha = 0.05$).

Bangladesh

Effect Size	Power	95% CI
0.005	19.50%	(17.09, 22.09)
0.0075	37.00%	(34.00, 40.08)
0.010	56.30%	(53.16, 59.40)
0.012	76.10%	(73.33, 78.71)
0.015	88.80%	(86.68, 90.69)
0.018	**95.90%**	**(94.48, 97.04)**
0.020	99.30%	(98.56, 99.72)

Indonesia

Effect Size	Power	95% CI
0.005	7.40%	(5.85, 9.20)
0.010	15.10%	(12.94, 17.47)
0.015	30.30%	(27.46, 33.25)
0.020	46.20%	(43.08, 49.35)
0.025	64.80%	(61.75, 67.76)
0.030	81.40%	(78.85, 83.77)
0.032	84.00%	(81.58, 86.22)
0.034	86.90%	(84.65, 88.93)
0.036	91.70%	(89.81, 93.34)
0.038	93.90%	(92.23, 95.30)
0.040	**96.10%**	**(94.71, 97.21)**

Power of the current study to detect a significant independent effect of Social Distance on Expected Sharing, as a function of varying sample sizes. Analysis uses the powerCurve() function from the SIMR package in R, which runs powerSim() over a range of sample sizes. This allows estimation of the number of participants necessary to have sufficient power to detect an effect of the size estimated in the model. All analyses use the same specifications described for powerSim() above.

Indonesia

Power to detect an independent effect [$\beta = -0.006$] of social distance on generosity from the model in Table 1 (main text).

Sample Size	Power	95% CI
44	7.20%	(5.68, 8.98)
1144	45.40%	(42.28, 48.55)
2224	72.60%	(69.72, 75.34)
3344	88.10%	(85.93, 90.04)
4224	**95.00%**	**(93.46, 96.27)**

Bangladesh

Power to detect an independent effect [$\beta = 0.002$] of social distance on generosity from the model in Table 1 (main text).

Sample Size	Power	95% CI
200	11.30%	(9.40, 13.43)
1000	49.50%	(46.36, 52.65)
1800	77.00%	(74.26, 79.58)
2600	89.40%	(87.32, 91.24)
2800	90.00%	(89.05, 92.70)
3200	**94.70%**	**(93.12, 96.01)**

INCONSISTENT RESPONDING ACROSS SITES

We found high rates of inconsistency (i.e. multiple crossover points for at least 1 recipient) among participants in both Bangladesh and Indonesia. When considering all participants and social distances (i.e. #1, #2, #5, #10, and #20), 28 out of 200 Bangladesh participants, 42 out of 44 Indonesia participants, and 9 out of 40 U.S. were inconsistent. This underestimates rates of inconsistency in Bangladesh, since 165 out of 200 participants always chose the selfish option and were considered consistent as a result. When considering participants with non-zero generosity (i.e. those who chose the generous option at least once for at least one recipient at social distances 1 to 20), 80% Bangladesh participants (28/35) and 100% of Indonesia participants (42/42) had at least 1 inconsistent response, compared to only 26% of U.S. participants (9/39). These are strikingly high levels of inconsistency. Figure 1S plots levels of inconsistency in these 3 sites alongside all reported inconsistency rates in social-discounting studies citing Rachlin and Jones' seminal paper (Jones & Rachlin, 2006) and using a comparable protocol (data and inclusion criteria: https://osf.io/k8sbg/). In contrast to U.S. participants and the vast majority of previous studies, inconsistency is the norm among the participants from rural Indonesia and Bangladesh.

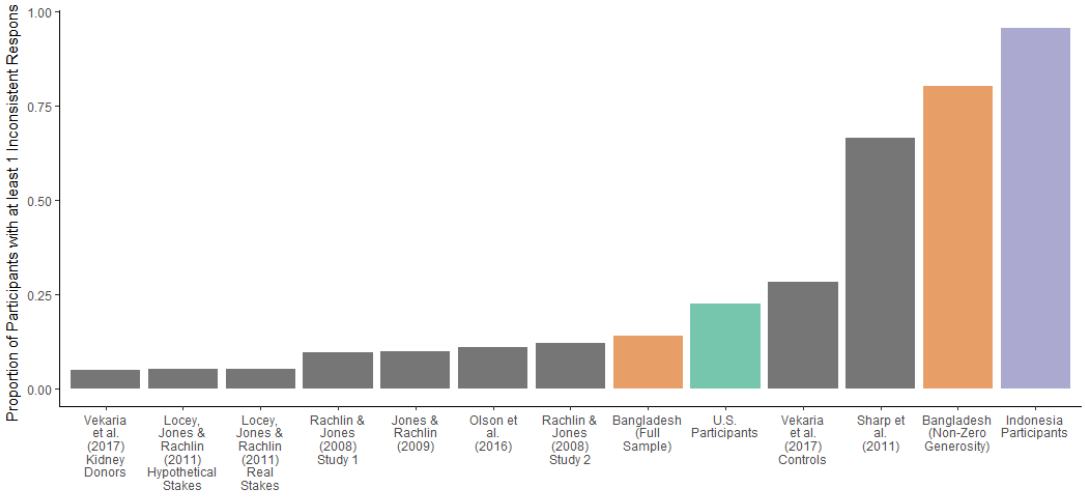


Figure 1S| Inconsistent responding. Proportion of inconsistent participants (i.e. multiple crossover points for at least 1 recipient) in prior social-discounting research, compared with levels of inconsistency in the U.S., Bangladesh, and Indonesia samples from the current study. Participants from Bangladesh are represented twice (including/excluding participants with non-zero generosity).

Several observations contradict the standard interpretation that "inconsistency" reflects lack of participant understanding. First, responses in all sites were associated with a theoretically important variable—whether the recipient was needier than the participant. Second, participants were more likely to choose the generous option when the payoff for the selfish option was small (See "Generosity as a Function of Payoff to Participant" in Supplementary Materials). Third, observations during piloting suggested that participants may not make (1) independent decisions based on (2) a constant utility function. For example, Bangladeshi participants often spoke out loud when making decisions. In such cases, many participants mentioned their decisions in previous choices while weighing current choices (e.g. "Well, I didn't give up 1kg in the last decision, so I'll give up 2kg this time."; "I already gave up 5kg and 2kg rice, so I won't give up 3kg this time"). This suggests that participants treated these as aggregate contributions, rather than as independent decisions. From this perspective, "inconsistent" responding with multiple crossover points is completely reasonable, and suggests the common model used to interpret consistent responding is wrong, at least in some situations.

EXPECTED SHARING AS A MEASURE OF GENEROSITY

"Expected sharing" is the weighted sum of all generous decisions divided by the weighted sum of all possible decisions (I.e. 0.5*X0 + 1*X1 + 2*X2 + 3*X3 + 4*X4 + 5*X5)/15.5. Here Xi = 1 if a participant sacrificed i units to give person 5 units. X0 indicates a sacrifice of 0.5 units. Although not identical to crossover points, expected sharing is monotonically increasing, provides a simple measure of average generosity towards a specific individual, and does not force exclusion of inconsistent respondents. Consider a participant who chooses to transfer $5 to the recipient, keeps $4, $3, $2, for themselves, and also chooses to transfer $1 and $0.50 to the recipient. In this case, expected sharing is simply (5 + 1 + 0.5) / (5 + 4 + 3 + 2 + 1 + 0.5) = 0.419.

RELATIONSHIP BETWEEN EXPECTED SHARING AND CROSSOVER POINTS

Typical analyses calculate the "crossover point" in the sequence of questions where respondents switch from the selfish option to the generous option(Jones & Rachlin, 2006). Consider a consistent participant who chooses the selfish option at $5 and $4 but switches to the generous option for subsequent decisions (i.e. $3, $2, $1, $0.50). This participant's crossover point is $3.50. A participant who always chooses the generous option (i.e. chooses the generous option at $5, $4, $3, $2, $1, $0.50) is considered to have a crossover point of $5.50. A participant who always chooses the selfish option (i.e. chooses the selfish option $5, $4, $3, $2, $1, $0.50) is considered to have a crossover point of $0.25. Any given crossover point corresponds to an exact expected sharing value. For our study, these values are:

Crossover Point	Expected Sharing
0.25	0
0.75	0.5 / 15.5
2.5	1.5 / 15.5
2.5	3.5 / 15.5
3.5	6.5 / 15.5
4.5	10.5 / 15.5
5.5	1

To approximate crossover points from expected sharing values for all participants, we fit a quadratic function to these values, of the form y = 0.0001669 + 0.0149876x + 0.0302354x^2. We then used the approx() function in R(R Core Team, 2017) (https://stat.ethz.ch/R-manual/R-devel/library/stats/html/approxfun.html) to perform linear interpolation, calculating approximate crossover points for every expected sharing value. Figure 2S (below) below plots the relationship between expected sharing and crossover points.

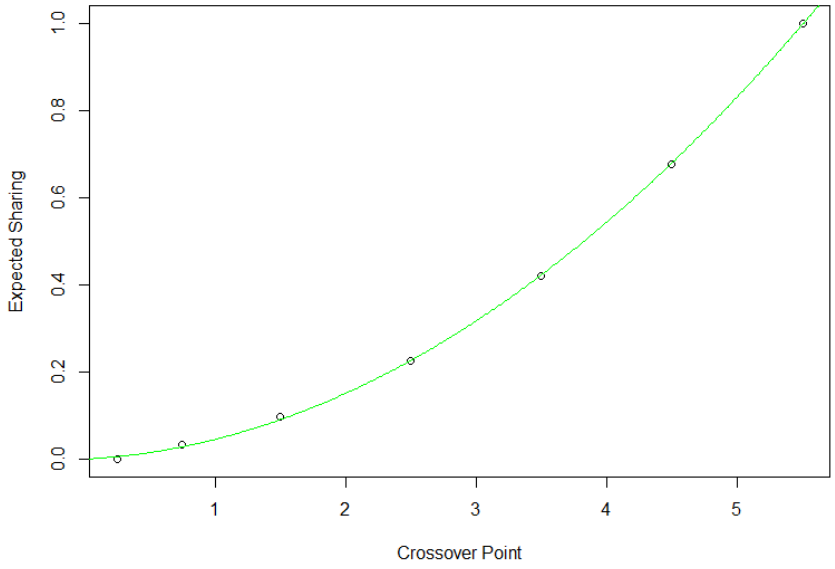


Figure 2S | Relationship Between Expected Sharing and Crossover Points.

RE-ANALYSIS USING APPROXIMATE CROSSOVER POINTS

	U.S. Crossover Points		Bangladesh Crossover Points		Indonesia Crossover Points	
	Estimate (CI)	*P*	*Estimate (CI)*	*P*	*Estimate (CI)*	*P*
Fixed Effects						
(Intercept)	4.32 (3.64, 4.99)	<.001	1.16 (0.76, 1.55)	<.001	4.22 (3.65, 4.79)	<.001
Ln Social Distance	-0.55 (-0.66, -0.44)	<.001	0.01 (-0.04, 0.06)	.775	-0.03 (-0.15, 0.08)	.545
Need						
Recipient Equally Needy	-0.29 (-0.91, 0.33)	.370	-0.36 (-0.77, 0.05)	.090	-0.77 (-1.27, -0.28)	.003
Recipient Less Needy	-0.80 (-1.45, -0.15)	.018	-0.75 (-1.15, -0.36)	<.001	-1.45 (-2.01, -0.89)	<.001
Relatedness	0.44 (-0.28, 1.16)	.232	-0.07 (-0.55, 0.40)	.759	0.60 (-0.04, 1.25)	.066
Random Parts						
σ^2	0.403					
$\tau_{00,\ respid}$	2.192					
ρ_{01}	-0.696					
N_{respid}	284					
Observations	1388					
R^2 / Ω_0^2	.916 / .914				.	

Table 20S| Generosity as a function of social distance, relative need, and relatedness, using approximated crossover points instead of expected sharing. Multilevel model of social distance, recipient need, and relatedness regressed on approximate crossover points. Model controls for correlated observations from the same participant with random effects for each individual and includes random slopes for social distance and recipient need. CI = 95% confidence intervals.

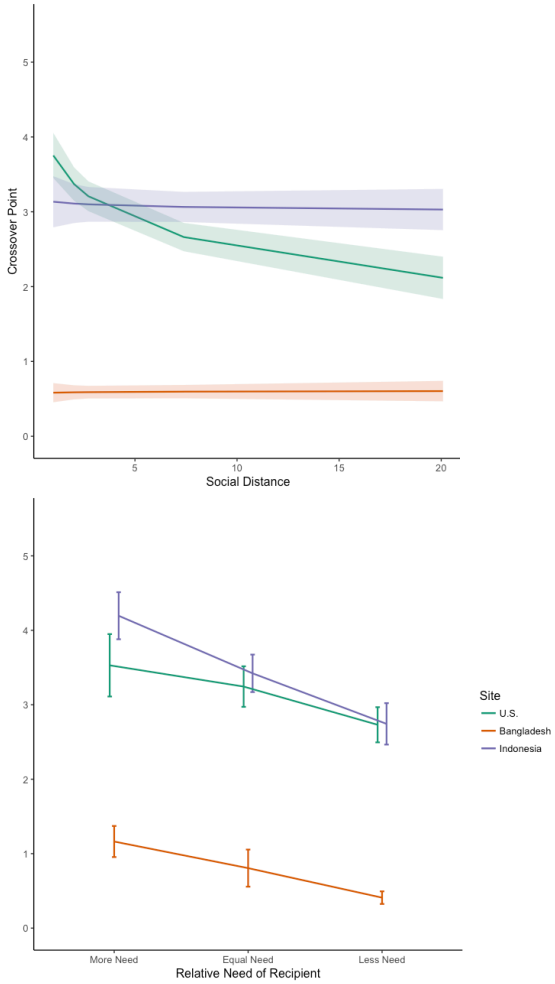


Figure 3S| Generosity as a function of social distance, relative need, and relatedness, using approximated crossover points instead of expected sharing. Independent effects of **a.** social distance (natural log transformed) and **b.** relative need on generosity (i.e. approximated crossover points). Model estimates from the multilevel model in Table 17S. Error bars represent 95% confidence intervals.

CODEBOOK FOR VERBAL STATEMENTS

Need. Any mention of the respondent's or recipient's need, financial situation, or wealth as a reason for the decision, with implication that either the respondent or recipient is in greater need. Includes mentions of general statements such as "it is good to help people who need money" and includes mentions of giving to someone who is in a needy situation (e.g. unemployed; widowed). Also includes mentions of pitying the recipient.
Relationship. Any mention of closeness, love, staying in touch, family, length of relationship or being in a relationship (e.g. good friend) as a reason for the decision. (Example: "He is my brother, but he is better off than me" would be coded as relationship, in addition to need). Includes mentions of living near the person and mentions of not knowing the person as a reason.

> **Relationship Exclusions:** If the relationship term is simply used as a description of the person (Example: "my brother is better off than me" would not be coded as relationship). If the only mention of the relationship is when respondent states that the recipient will later use the money on them or that the respondent owes the recipient. If the only reason for mentioning relationship is not knowing the other person (Example: "This person is a stranger" or "I don't know this person"), without an implication that the respondent made their decision because they did not know this person, this would not be coded as "relationship".

Reciprocity / Imbalance. Mentions giving because the other person has given to them in the past, giving in order to have the other person give to them in the future, mentions "establishing a reciprocal relationship" as a good thing, or mentions "taking turns". Also includes references to the fact that the respondent owes the recipient or that the recipient has already received too much from the respondent.
Moral. Reasoning that it is good to help others.
Descriptive. Simply states what they did (e.g., I gave to myself) without any clear reason given.
Efficiency. Refers to the fact that the respondent would give up less than the other person got.
Make Happy. Refers to the feeling that might be invoked in the recipient by the gift (happy, excited, etc.). Includes instances where respondents mentioned that the recipient would "appreciate" the gift.
Indirect Benefit. If the respondent states that the recipient will later use the money on them.
Give to something else. The respondent will keep the money to give to something or somebody else later or do something for the recipient later.
Previous decision. Respondent describes earlier decision as a justification for a later decision (I already sacrificed, so I kept for myself later; I already kept for myself, so I decided to give).
Want the money. Respondent simply stated they wanted the money.
Deserving. Respondent or recipient deserves or doesn't deserve it for reasons other than need (e.g. person is hard working, profligate)

Other. Anything that doesn't fall into the other categories (e.g. mentions of the appropriateness of an action, mentions of religion, etc.). Reasoning that it is good to help others

APPENDIX B

CHAPTER 2

Model

To calculate a player's expected payoff for guessing the majority color, we simulated the average amount of information available to a player, conditional on having revealed a given number of tiles (see "Model" in main text). Below, Figure 1S plots information as a function of number of tiles revealed, for the 3 effect sizes used in the experiment.

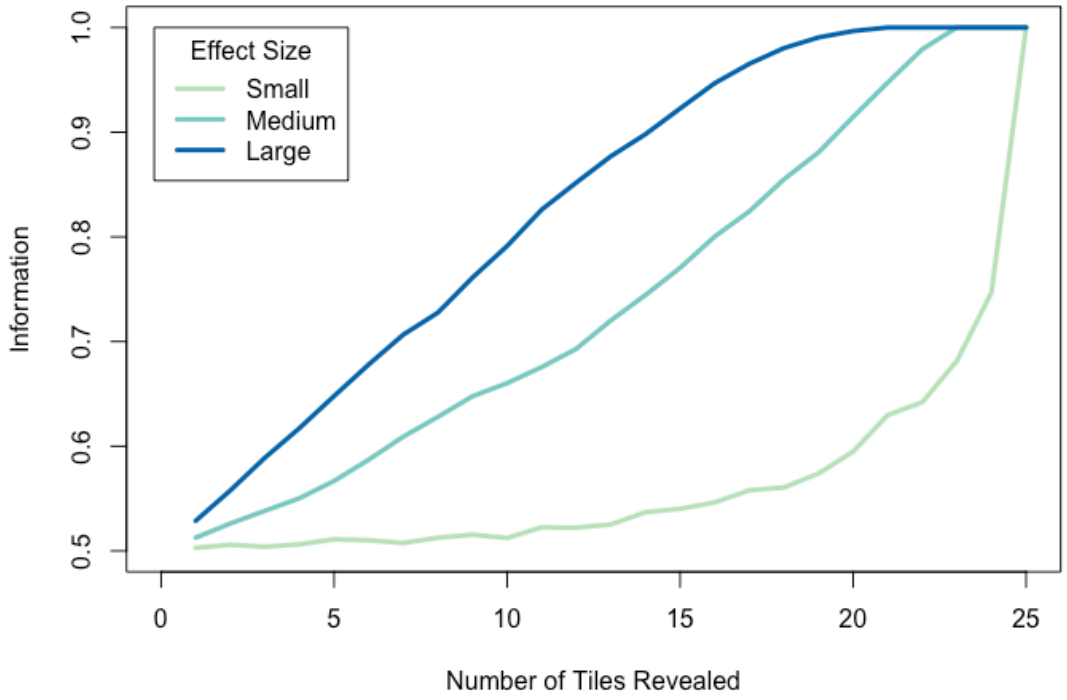


Figure 1S| Average information (500 simulations) as a function of the number of tiles revealed and effect size. Plotted for the three ratios of colored tiles used in the experiment (i.e. effect sizes). "Small", "Medium", and "Large" effect sizes correspond to colored-tile ratios of 12:13, 10:15, and 8:17, respectively. The X axis indicates the number of tiles revealed. The Y axis indicates the information available to players, averaged across 500 simulations. In the model, we assume that players guess the majority color by selecting the color that is in the majority in the tiles that they reveal. Information thus corresponds to the probability of correctly guessing the majority color (p_o and p_p). For any given number of tiles revealed, information is larger when effect sizes are larger.

Given the information value of revealing any given number of tiles, we can calculate a player's expected number of points (i.e. reward, R) by the completion of the experiment in the No-Competition, No-Effort treatment. To calculate R for each effect size, we assume that an individual plays a 20-minute (1200 second) experiment with grids that are characterized by a single effect size.

In this treatment, players gain or lose 1 point by guessing the majority color correctly or incorrectly, respectively. Players can reveal 1 tile every 1 second, and experience a 5-second delay between guessing the majority color of a grid and being able to start another grid. As such, a player's expected reward, *R*, is:

$$R = \frac{1200}{n} * (p - (1 - p)) * \frac{n}{n + 5}$$
$$R = \frac{2400p - 1200}{n + 5}$$

where *p* is the probability that the player guesses correctly and *n* is the number of tiles they reveal. 1200 corresponds to the length of the experiment (seconds) and 5 corresponds to the 5-second delay between grids. Players who reveal fewer tiles have the opportunity to guess the majority color more often: they encounter more grids within the 20-minute time limit. However, those players also have a lower probability of correctly guessing the majority color, and experience the 5-second delay between-grids more often. The number of tiles that maximizes a player's expected reward in the No-Competition No-Effort treatment is 25, 23, and 16, for small, medium and large effect sizes respectively. Below, Figure S2 plots player's expected reward as a function of number of tiles revealed, for the 3 effect sizes used in the experiment.

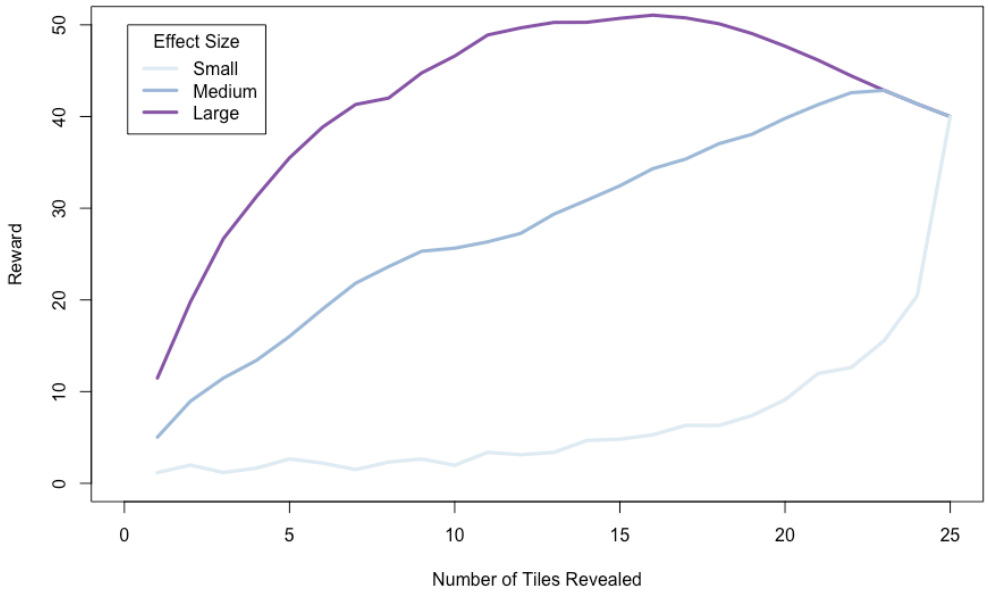


Figure 2S| Expected reward as a function of the number of tiles revealed and effect size, without competition. Plotted for the three ratios of colored tiles used in the experiment (i.e. effect sizes). "Small", "Medium", and "Large" effect sizes correspond to colored-tile ratios of 12:13, 10:15, and 8:17, respectively. The X axis indicates the number of tiles revealed. The Y axis indicates the expected reward, for a player who reveals a fixed number of tiles. The number of tiles that

maximizes a player's expected reward in the No-Competition No-Effort condition, for small, medium and large effect sizes is 25, 23, and 16, respectively.

Assuming that a player in the Competition, No-Effort treatment competes against a competitor who revealed the number of tiles that maximized their expected reward, we can calculate a player's expected payoff (*EP)* for revealing any given number of tiles (as in the main text).

When a player guesses before or at the same time as an opponent, the player's expected payoff (*EP*) is:

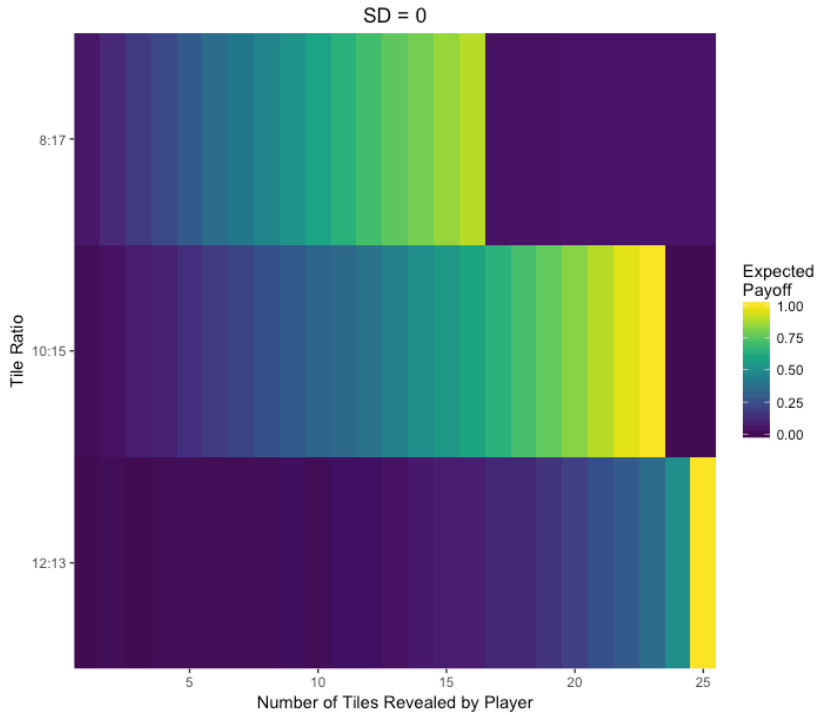
$$EP = p_p - (1 - p_p)$$

where p_p is the probability that the player guesses correctly. When a player guesses after their opponent, the player's *EP* is:

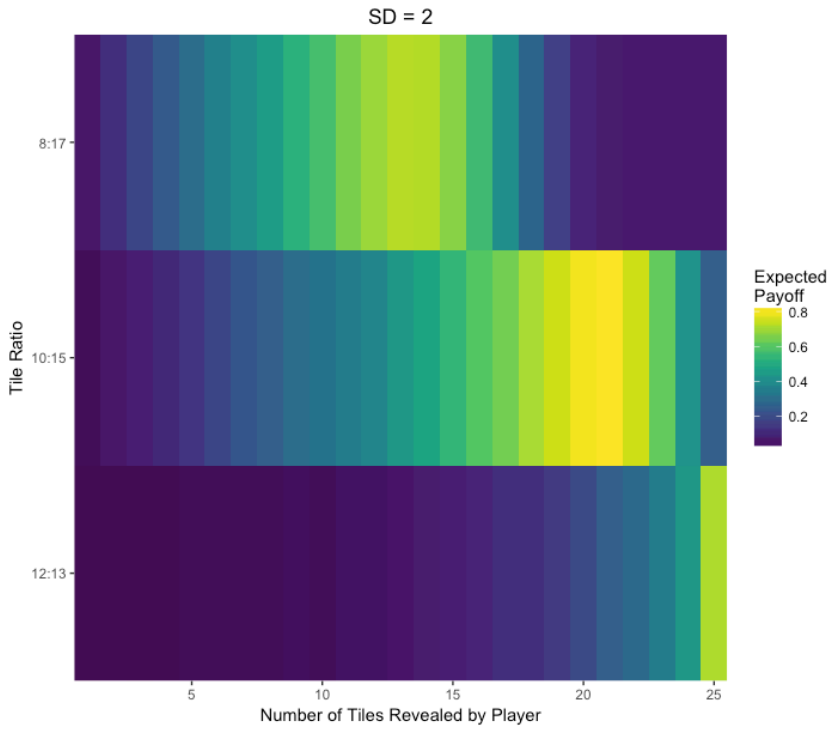
$$EP = (1 - p_o) * p_p - (1 - p_o) * (1 - p_p)$$

where p_o is the probability that the player's opponent (i.e. the player in the no-competition treatment) guesses correctly. In this case, by assuming that opponents reveal 25, 23, and 16 tiles for small, medium, and large effect sizes, respectively, we know the amount of information available to players who reveal any possible number of tiles (Figure S2), and can directly calculate *EP*. Figure S3a (below) depicts a player's *EP* when competing against a competitor who always reveals the payoff-maximizing number of tiles. Figures S3b and S3c relax the assumption that a competitor always reveals the payoff-maximizing number of tiles by assuming that the number of tiles that a competitor reveals is a rounded value sampled from a normal distribution with a mean equal to the payoff-maximizing number of tiles, and standard deviation of 2 and 5, respectively. Values < 1 and > 25 are rounded to 1 and 25, respectively. For Figures S3b and S3c, player's *EP* is calculated by averaging across 10,0000 samples from these distributions.

a.


SD = 0
8:17
10:15
12:13
Tile Ratio
5
10
15
20
25
Number of Tiles Revealed by Player
Expected Payoff
1.00
0.75
0.50
0.25
0.00

b.


SD = 2
8:17
10:15
12:13
Tile Ratio
5
10
15
20
25
Number of Tiles Revealed by Player
Expected Payoff
0.8
0.6
0.4
0.2

c.

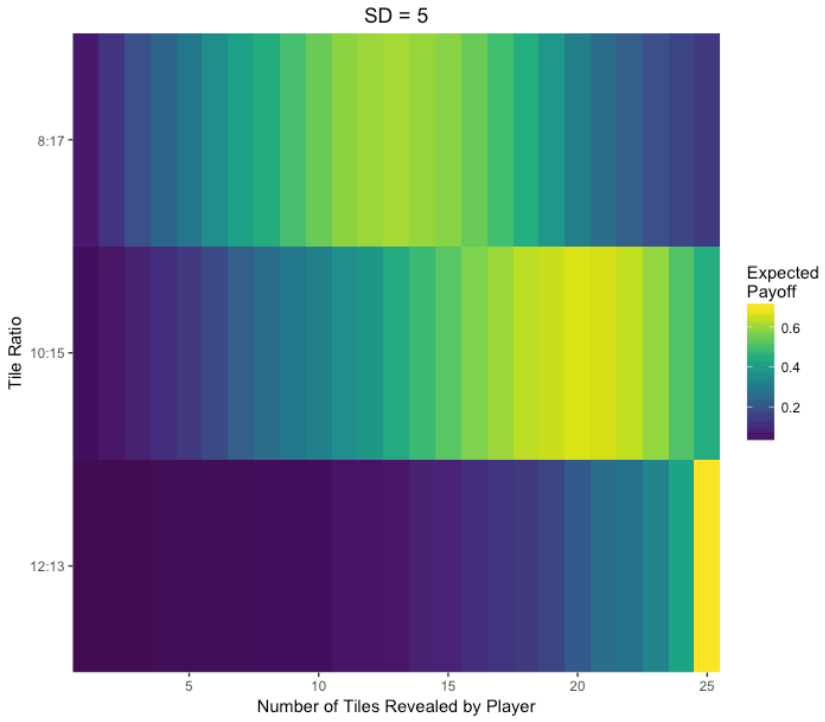


Figure 3S| Expected payoff as a function of the number of tiles revealed and effect size, when playing against a competitor who reveals the payoff-maximizing number of tiles. Plotted for the three ratios of colored tiles used in the experiment (i.e. effect sizes). "Small", "Medium", and "Large" effect sizes correspond to colored-tile ratios of 12:13, 10:15, and 8:17, respectively. The X axis indicates the number of tiles revealed by a player in the Competition treatment. **a.** Competitor always reveals the number of tiles that maximizes their expected payoff. **b and c.** The number of tiles revealed by a competitor is a random variable, sampled from a normal distribution with a mean equal to the payoff-maximizing number of tiles (25, 23, and 16 for small, medium, and large effect sizes, respectively), and standard deviation of 2 and 5, respectively; player's *EP* is calculated by averaging across 10,0000 samples from these distributions. In general, players maximize their expected payoff by revealing the same number or fewer tiles than their competitor.

This analysis corroborates the intuition suggested by a visual inspection of Figure 2 in the main text: players obtain a higher *EP* by revealing the exact same or fewer tiles than their competitors. This is true when competitors always reveal the payoff-maximizing number of tiles, and when the number of tiles revealed by a competitor is a random variable instead of a fixed value.

Pilot Study

We conducted a pre-registered (https://osf.io/udm8g/) pilot study. This study was designed to test the feasibility of the proposed design, not to test hypotheses. In conducting the pilot study, we underspecified exclusion criteria and deviated in several ways from pre-specified pilot analysis plan. As such, we consider all findings from this pilot study to be exploratory.

The pilot study involved 48 participants and was conducted in the Elinor Ostrom Multi-Method Lab at Arizona State University. We excluded data from 1 participant that did not complete the study, resulting in a final sample of 47 participants (23 female, 24 male). 16 and 31 participants were assigned to the Competition and No-Competition treatments and 23 and 24 participants were assigned to the Effort and No-Effort conditions, respectively. The pilot study differed from the proposed design in one way:

1) Players were paid \$0.25 cents per solution instead of \$0.15 cents:

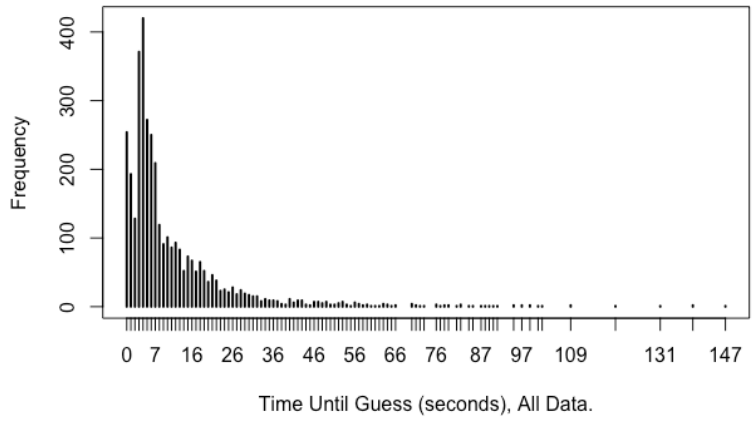
   $Payoff_{no\text{-}competition}$ = \$0.25 x CorrectGuesses – \$0.25 x IncorrectGuesses
   $Payoff_{competition}$ = \$0.25 x CorrectFasterGuesses – \$0.25 x IncorrectFasterGuesses

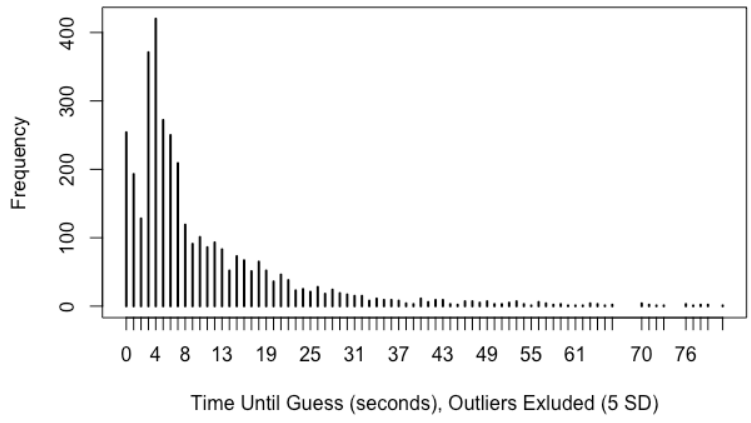
Below, we present the results for quality checks and all confirmatory predictions. We present three pieces of information for each analysis: 1) parameter estimates from the proposed Bayesian statistical model, 2) parameter estimates from Frequentist implementations of the same model, and 3) a plot visualizing the predictions of the Frequentist-model. Because some pilot analyses differ from the previously proposed analyses, we specify the statistical model for each analysis.

Exclusions and Outliers.

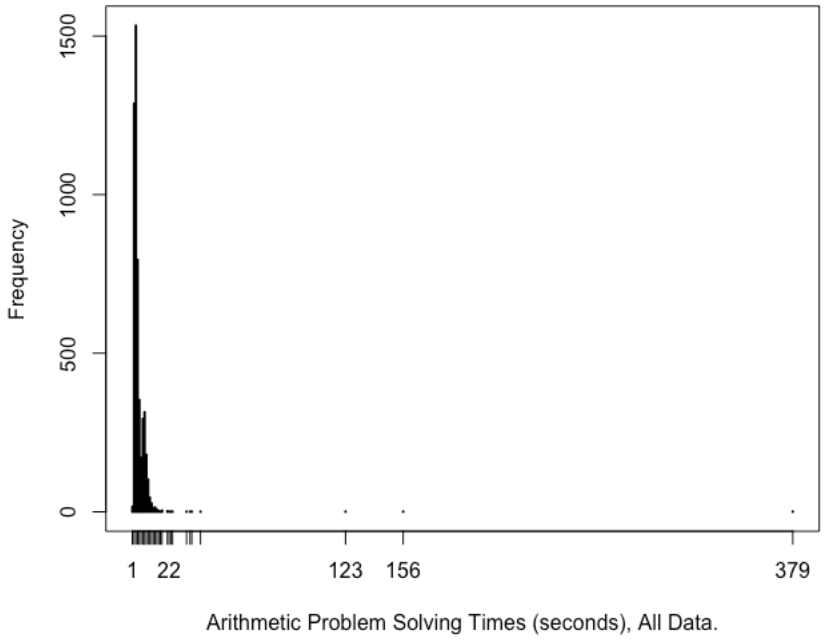
Both participants' time until guess and time to solve arithmetic problems followed heavily right-skewed distributions (see below). We removed outlier time-until-guess values that were more than 5 standard deviations larger than the mean time-until-guess (25 out of 3565 observations). This resulted in excluding substantially fewer observations than would be excluded had we used an outlier criterion of larger than 3 standard deviations (79 observations). We removed outlier arithmetic-problem solving times that were more than 5 standard deviations larger than the mean arithmetic-problem solving time (4 out of 5169 observations). This resulted in excluding slightly fewer observations than would be excluded had we used an outlier criterion of larger than 3 standard deviations (8 observations). Below, we present visualizations of the time until guess and arithmetic-problem solving times (both rounded to the nearest second) before and after outlier exclusion.

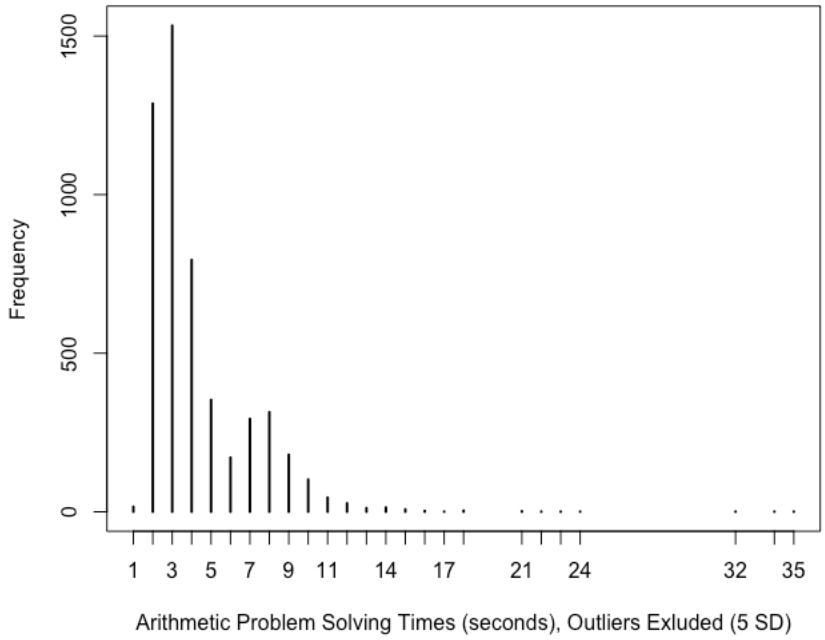
Time Until Guess





Arithmetic Problem Solving Times





Quality Checks.

1. Participants should reveal information at a lower rate in the Effort treatments than the No-Effort treatments.

Because we did not obtain data on time-per-click in the no-effort treatment, we cannot use the pilot data to perform this quality check. A complementary quality check is that participants in the Effort treatments should spent more time per grid (i.e. take longer to guess the underlying color) than participants in the No-Effort treatment. This quality check was confirmed (see Exploratory Analysis 2, "Time Until Guess", below).

2. A higher proportion of participants in the Competition treatments should answer "yes" to a question about whether or not they were competing with another player.

This quality check was confirmed. 16/16 participants answered "yes" in the Competition treatments, compared to 3/31 participants in the No-Competition treatments. Below, we present parameter estimates from the Bayesian model only (the frequentist implementation did not converge because "yes" responses were almost completely separated by the predictor).

Likelihood

$$Y_i \sim \text{Binomial}(1, p_i)$$

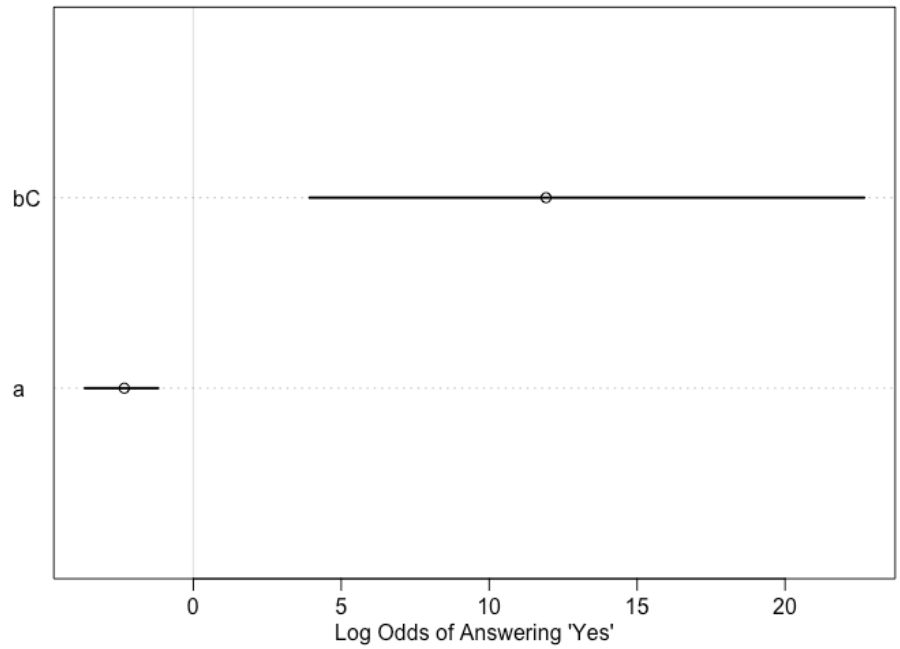
$$\text{Logit}(p_i) = \alpha + \beta_C * C_i$$

Y_i: Answered "yes". α: Intercept. C: Competition Treatment (1 / 0).

Priors

$$\alpha \sim \text{Normal}(0, 10)$$
$$\beta_C \sim \text{Normal}(0, 10)$$

Results



Exploratory Analyses.

Competition and number of tiles revealed (see H1a and H3a).

Likelihood

$$Yi \sim \text{Normal}(\mu_i, \sigma)$$

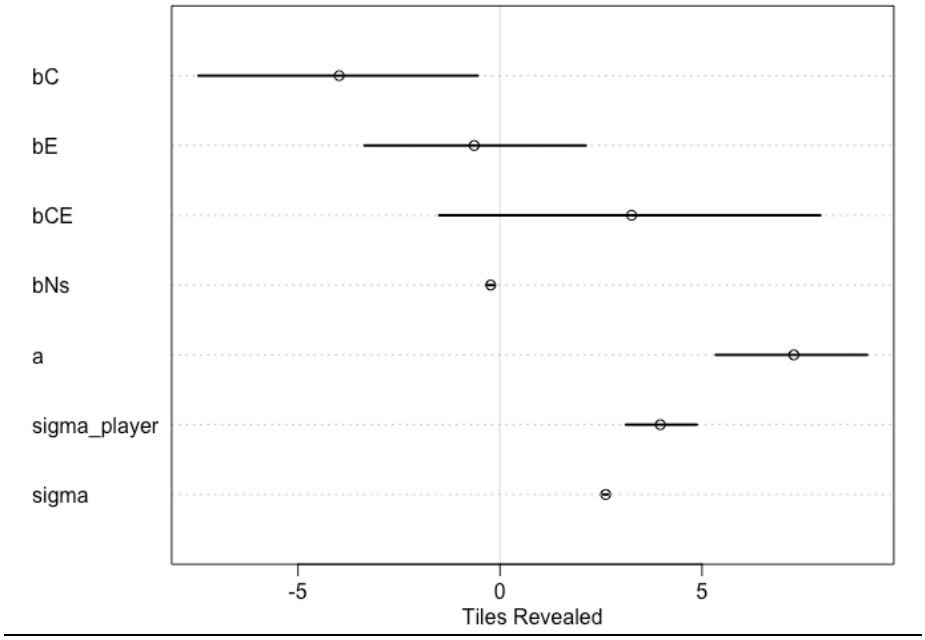
$$\mu_i = \alpha + \alpha_{\text{PLAYER}[i]} + \beta_C * C_i + \beta_E * E_i + \beta_{CE} * C_i E_i + \beta_{Ns} * Ns_i$$

Y_i: Number of tiles clicked before guessing. α: Intercept. $\alpha_{\text{PLAYER}[i]}$: Random intercept for each player. C: Competition Treatment (1 / 0). E: Effort Condition (1 / 0). $\beta_{CE} * C_i E_i$: Interaction between treatment and effort. β_{Ns}: Standardized number of tiles for the majority color (i.e. effect size).

Priors

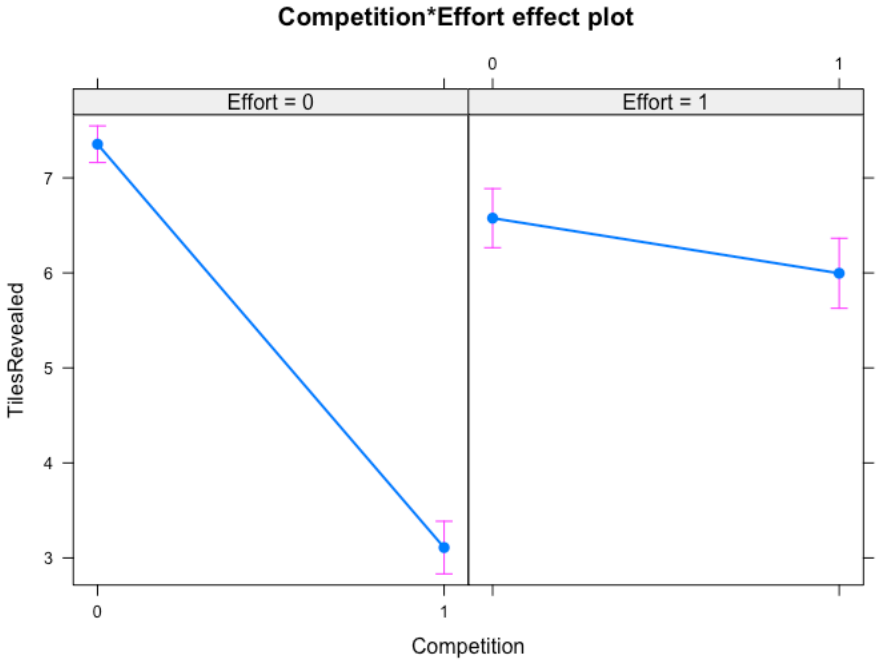
$$\sigma \sim \text{Gamma}(2, 0.5)$$
$$\alpha \sim \text{Uniform}(0, 25)$$
$$\alpha_{\text{PLAYER}} \sim \text{Normal}(0, \sigma_{\text{PLAYER}})$$
$$\sigma_{\text{PLAYER}} \sim \text{Gamma}(1.5, 0.05)$$
$$\beta_C \sim \text{Normal}(0, 10)$$
$$\beta_E \sim \text{Normal}(0, 10)$$
$$\beta_{CE} \sim \text{Normal}(0, 10)$$
$$\beta_{Ns} \sim \text{Normal}(0, 10)$$

Tiles Revealed (Bayesian)



Tiles Revealed (Frequentist)

	TilesRevealed	
	Estimate (CI)	*P*
Fixed Parts		
(Intercept)	7.69 (5.94 – 9.45)	<.001
Competition1	-4.25 (-7.49 – -1.01)	.010
Effort1	-0.78 (-3.39 – 1.84)	.559
SmallEffect.s		
MediumEffect.s	-0.37 (-0.58 – -0.16)	<.001
LargeEffect.s	-0.61 (-0.83 – -0.38)	<.001
Competition1:Effort1	3.67 (-0.82 – 8.15)	.109
Random Parts		
σ^2	6.840	
$\tau_{00,\ ID\_Player}$	13.476	
N_{ID_Player}	47	
Observations	3540	
R^2 / Ω_0^2	.614 / .614	



Competition and accuracy (see H1b and H3b).

Likelihood

$$S_i \sim \text{Binomial}(1, p_i)$$

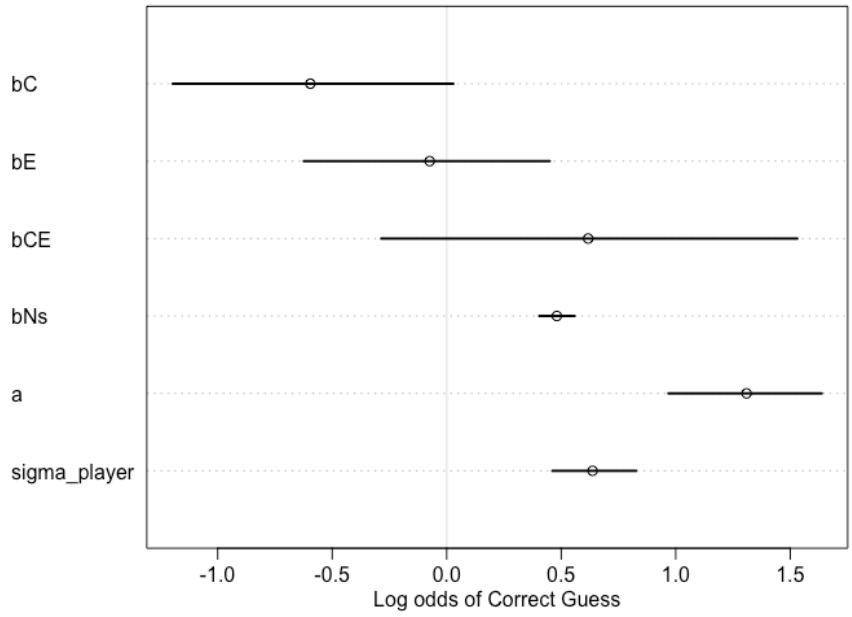
$$\text{Logit}(p_i) = \alpha + \alpha_{\text{PLAYER}[i]} + \beta_C * C_i + \beta_E * E_i + \beta_{CE} * C_i E_i + \beta_{Ns} * Ns_i$$

S_i: Successful guess. α: Intercept. $\alpha_{\text{PLAYER}[i]}$: Random intercept for each player. C: Competition Treatment (1 / 0). E: Effort Condition (1 / 0). $\beta_{CE} * C_i E_i$: Interaction between treatment and effort. β_{Ns}: Standardized number of tiles for the majority color (i.e. effect size).

Priors

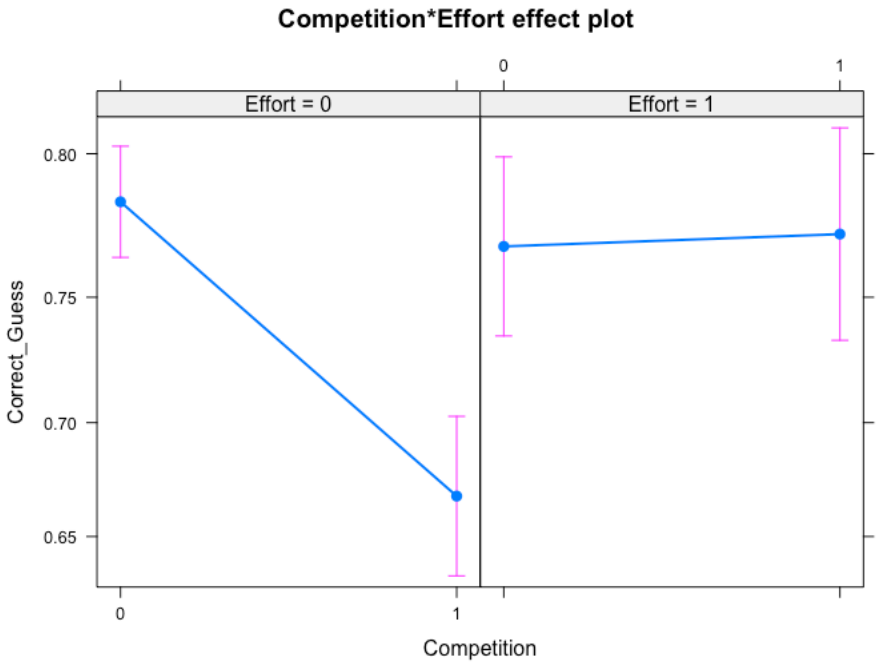
$$\sigma \sim \text{Gamma}(2, 0.5)$$
$$\alpha \sim \text{Normal}(0, 10)$$
$$\alpha_{\text{PLAYER}} \sim \text{Normal}(0, \sigma_{\text{PLAYER}})$$
$$\sigma_{\text{PLAYER}} \sim \text{Gamma}(1.5, 0.05)$$
$$\beta_C \sim \text{Normal}(0, 10)$$
$$\beta_E \sim \text{Normal}(0, 10)$$
$$\beta_{CE} \sim \text{Normal}(0, 10)$$
$$\beta_{Ns} \sim \text{Normal}(0, 10)$$

Probability of Correct Guess (Bayesian)



Probability of Correct Guess (Frequentist)

	Log Odds Correct_Guess	
	Estimate (CI)	*P*
Fixed Parts		
(Intercept)	0.58 (0.26 – 0.90)	<.001
Competition1	-0.59 (-1.13 – -0.05)	.032
Effort1	-0.09 (-0.57 – 0.39)	.716
SmallEffect.s		
MediumEffect.s	0.82 (0.64 – 0.99)	<.001
LargeEffect.s	1.23 (1.03 – 1.43)	<.001
Competition1:Effort1	0.61 (-0.19 – 1.42)	.133
Random Parts		
$\tau_{00,\ ID\_Player}$	0.318	
N_{ID_Player}	47	
Observations	3540	
Deviance	3955.318	



Competition and time until players guess

Likelihood

$$Yi \sim \text{Normal}(\mu_i, \sigma)$$

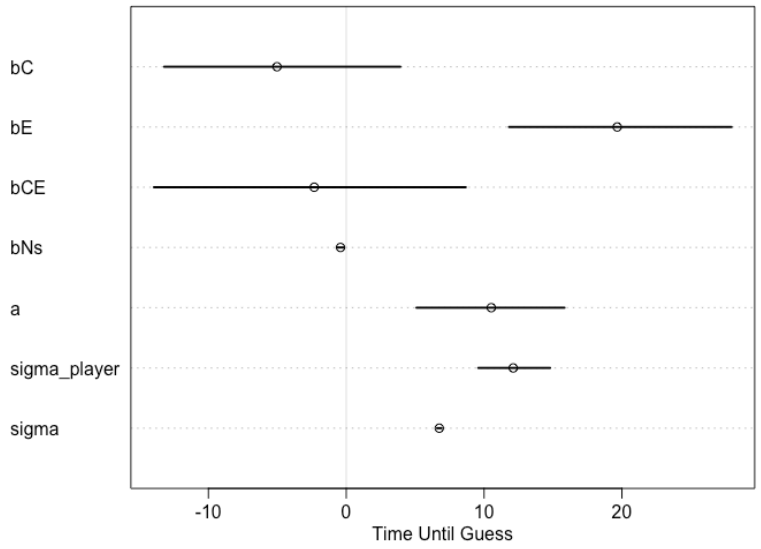
$$\mu_i = \alpha + \alpha_{\text{PLAYER}[i]} + \beta_C * C_i + \beta_E * E_i + \beta_{CE} * C_i E_i + \beta_{Ns} * Ns_i$$

Y_i: Number of seconds until guess. α: Intercept. $\alpha_{\text{PLAYER}[i]}$: Random intercept for each player. C: Competition Treatment (1 / 0). E: Effort Condition (1 / 0). $\beta_{CE} * C_i E_i$: Interaction between treatment and effort. β_{Ns}: Standardized number of tiles for the majority color (i.e. effect size).

Priors

$$\sigma \sim \text{Gamma}(2, 0.5)$$
$$\alpha \sim \text{Gamma}(1.5, 0.05)$$
$$\alpha_{\text{PLAYER}} \sim \text{Normal}(0, \sigma_{\text{PLAYER}})$$
$$\sigma_{\text{PLAYER}} \sim \text{Gamma}(1.5, 0.05)$$
$$\beta_C \sim \text{Normal}(0, 10)$$
$$\beta_E \sim \text{Normal}(0, 30)$$
$$\beta_{CE} \sim \text{Normal}(0, 10)$$
$$\beta_{Ns} \sim \text{Normal}(0, 10)$$

Time Until Guess (Bayesian)



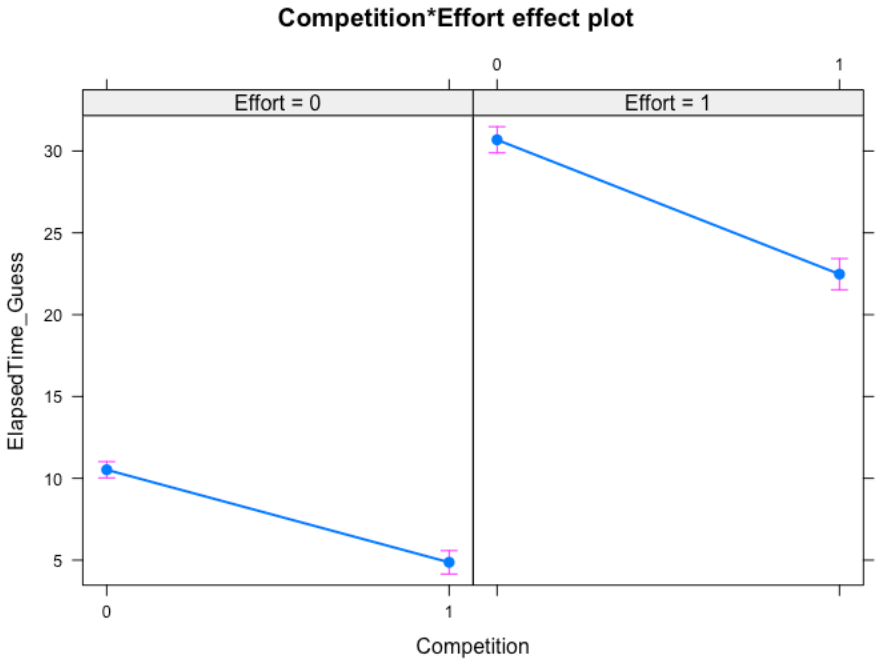
Time Until Guess (Frequentist)

	ElapsedTime_Guess	
	Estimate (CI)	*P*
Fixed Parts		
(Intercept)	11.15 (5.74 – 16.57)	<.001
Competition1	-5.66 (-15.65 – 4.34)	.267
Effort1	20.16 (12.11 – 28.22)	<.001
SmallEffect.s		
MediumEffect.s	-0.72 (-1.26 – -0.18)	.009
LargeEffect.s	-1.10 (-1.68 – -0.51)	<.001

Competition1:Effort1	-2.55 (-16.38 – 11.27)	.717

Random Parts

σ^2	45.519
$\tau_{00,\ ID\_Player}$	128.482
N_{ID_Player}	47
Observations	3540
R^2 / Ω_0^2	.668 / .668



Competition and number of guesses per unit time (i.e. guess rate)

Likelihood

$$Yi \sim \text{Normal}(\mu_i, \sigma)$$

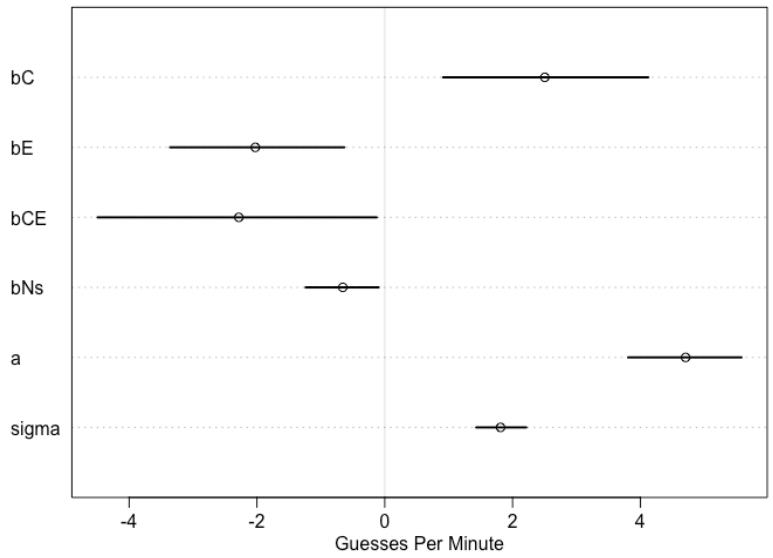
$$\mu_i = \alpha + \beta_C * C_i + \beta_E * E_i + \beta_{CE} * C_i E_i + \beta_{Ns} * Ns_i$$

Y_i: Guess Rate. α: Intercept. C: Competition Treatment (1 / 0). E: Effort Condition (1 / 0). $\beta_{CE} * C_i E_i$: Interaction between treatment and effort. β_{Ns}: Standardized mean number of tiles for the majority color (i.e. effect size) encountered by a player across all attempted grids.

Priors

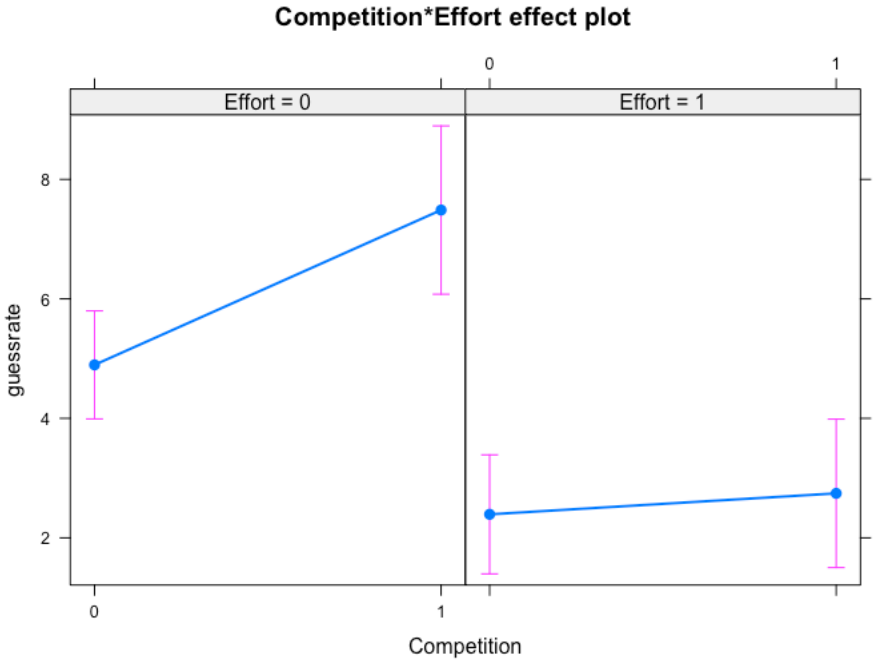
$$\sigma \sim \text{Gamma}(2, 0.5)$$
$$\alpha \sim \text{Gamma}(2, 0.05)$$
$$\beta_C \sim \text{Normal}(0, 10)$$
$$\beta_E \sim \text{Normal}(0, 10)$$
$$\beta_{CE} \sim \text{Normal}(0, 10)$$
$$\beta_{Ns} \sim \text{Normal}(0, 10)$$

Guess Rate (Bayesian)



Guess Rate (Frequentist)

	guessrate	
	Estimate (CI)	*P*
(Intercept)	4.68 (3.80 – 5.56)	<.001
Competition1	2.55 (0.96 – 4.14)	.002
Effort1	-1.99 (-3.34 – -0.64)	.005
MeanEffectSize.s	-0.66 (-1.22 – -0.10)	.022
Competition1:Effort1	-2.34 (-4.54 – -0.14)	.038
Observations	47	
R^2 / adj. R^2	.561 / .520	



Competition and effort (i.e. rate of revealing information; see H2).

Likelihood

$$Yi \sim \text{Normal}(\mu_i, \sigma)$$

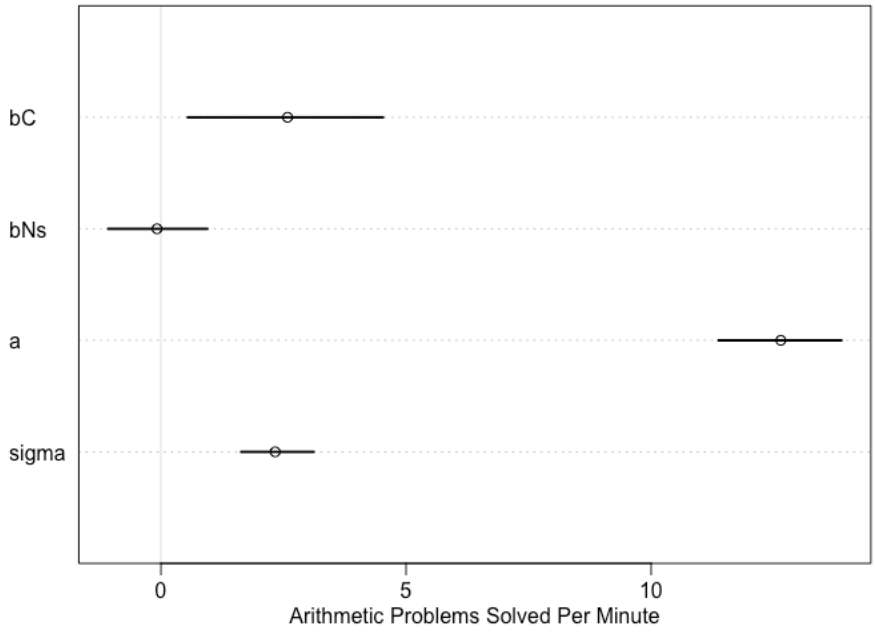
$$\mu_i = \alpha + \beta_C * C_i + \beta_{Ns} * Ns_i$$

Y_i: Rate of solving arithmetic problems. α: Intercept. C: Competition Treatment (1 / 0). β_{Ns}: Standardized mean number of tiles for the majority color (i.e. effect size) encountered by a player across all attempted grids.

Priors

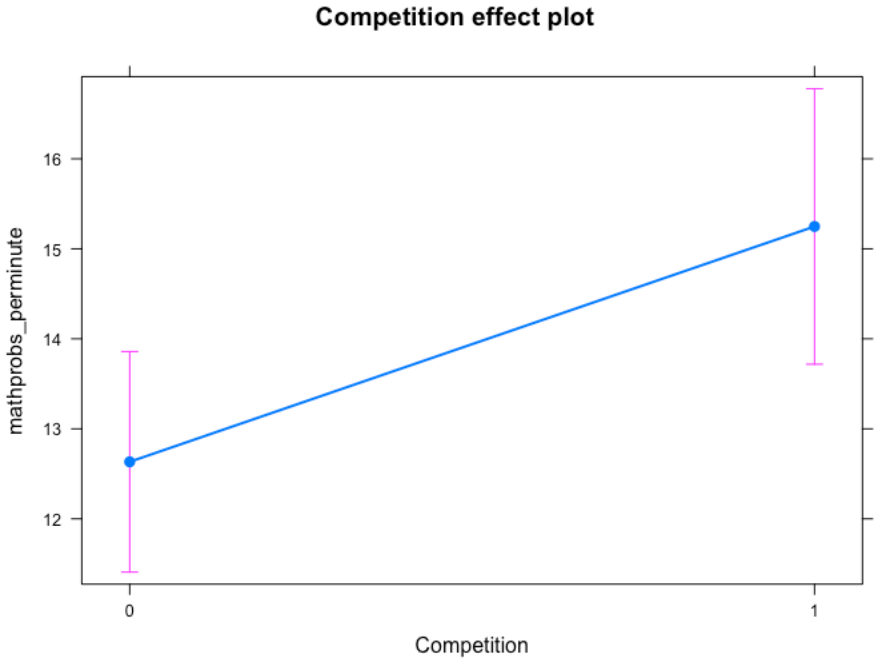
$$\sigma \sim \text{Gamma}(2, 0.5)$$
$$\alpha \sim \text{Gamma}(1.5, 0.05)$$
$$\beta_C \sim \text{Normal}(0, 10)$$
$$\beta_{Ns} \sim \text{Normal}(0, 10)$$

Rate of Solving Arithmetic Problems (Bayesian)



Rate of Solving Arithmetic Problems (Frequentist)

	mathprobs_perminute	
	Estimate (CI)	*P*
(Intercept)	12.63 (11.41 – 13.86)	<.001
Competition (1)	2.61 (0.65 – 4.58)	.012
MeanEffectSize.s	-0.08 (-1.07 – 0.90)	.861
Observations	23	
R^2 / adj. R^2	.285 / .213	
		183


Competition effect plot
mathprobs_perminute
16
15
14
13
12
0
1
Competition

APPENDIX C

CHAPTER 3

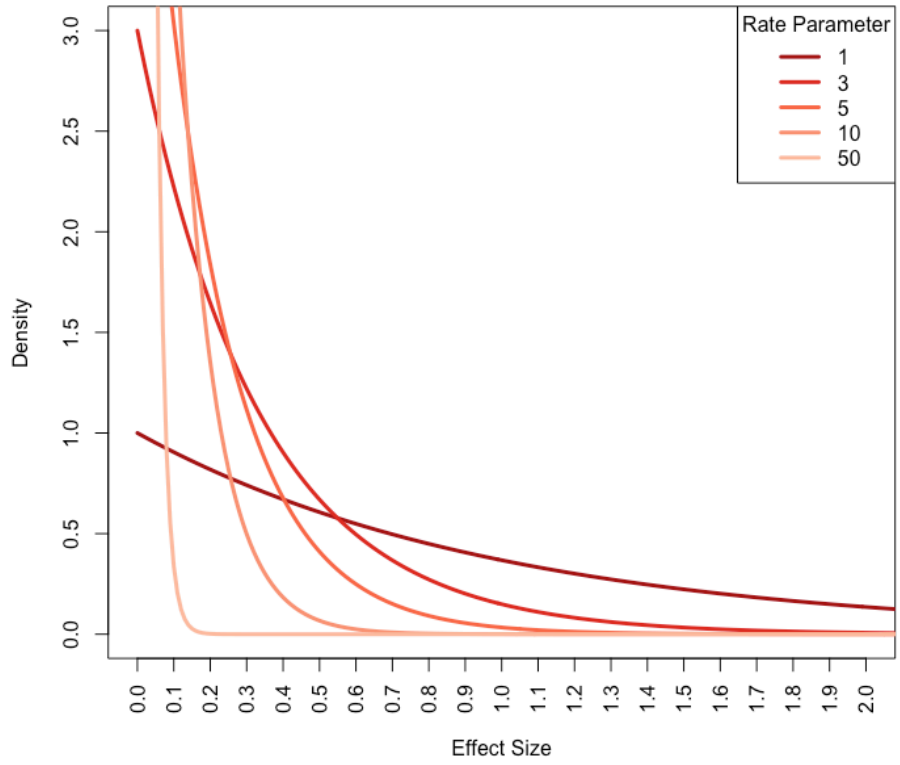
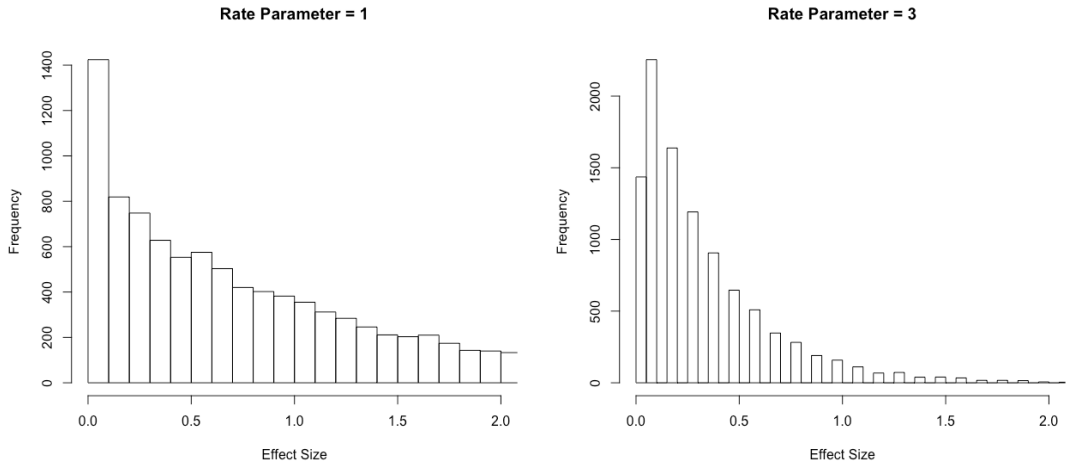


Figure 1S | Density of Effect Sizes as a function of different exponential-distribution rate (λ) parameters.



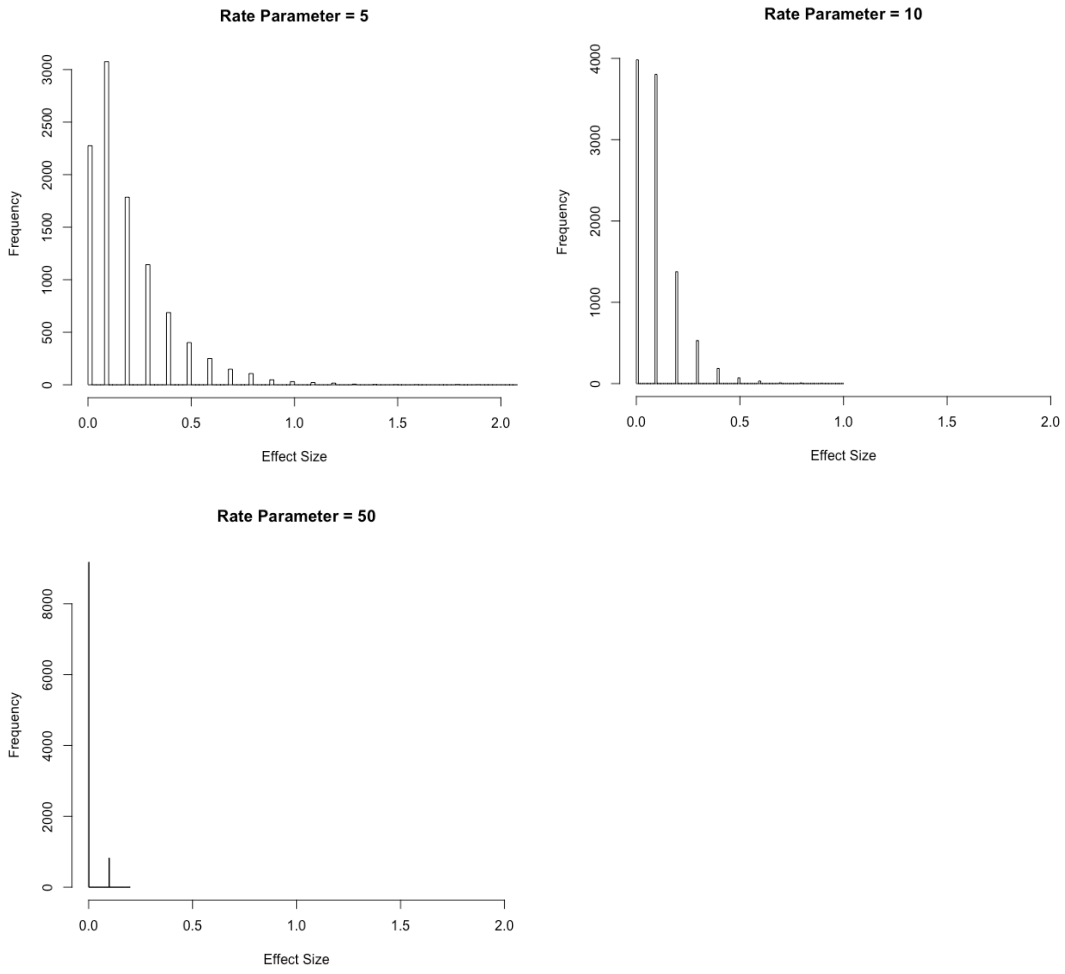


Figure 2S | Histogram of Rounded Effect Sizes as a function of different exponential-distribution rate (λ) parameters. Rounding without adding 0.1 to each effect size.

Fixed Effect Sizes (*e*) (i.e. not drawn from an exponential distribution)

a. **b.**

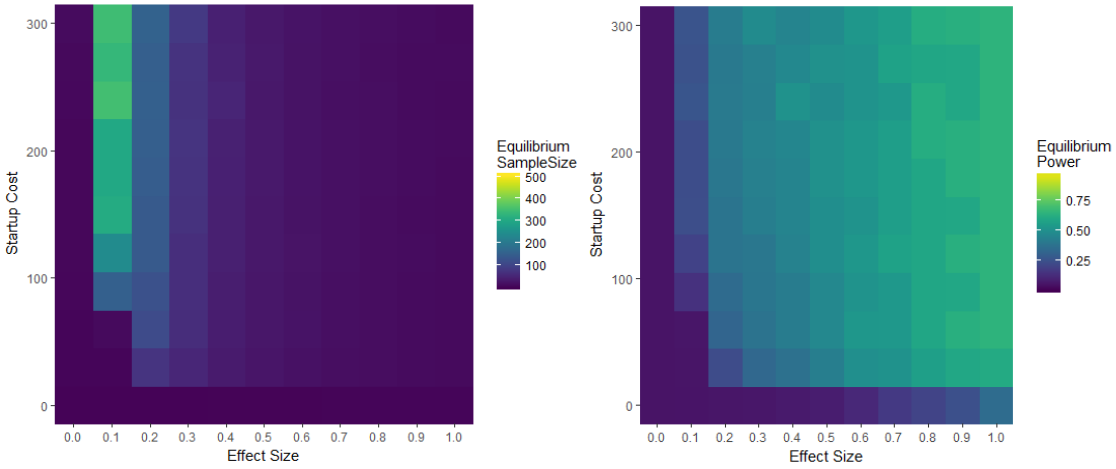


Figure 3S | Equilibrium (a) sample size and (b) statistical power as a function of effect size and startup cost (50 generations); 2 competitors. Parameter values are: n=100, α=0.05, r=5, T=5000 and c_s=1. a) Larger effect sizes lead to smaller sample sizes at equilibrium, while larger startup costs lead to larger sample sizes at equilibrium. When effect size = 0 (i.e. no effect), equilibrium sample sizes are at their lowest value. Equilibrium sample sizes are greatest when the effect size is small, but non-zero, and the startup cost is high. b) Statistical power at equilibrium is highest when both effect sizes and startup costs are large.

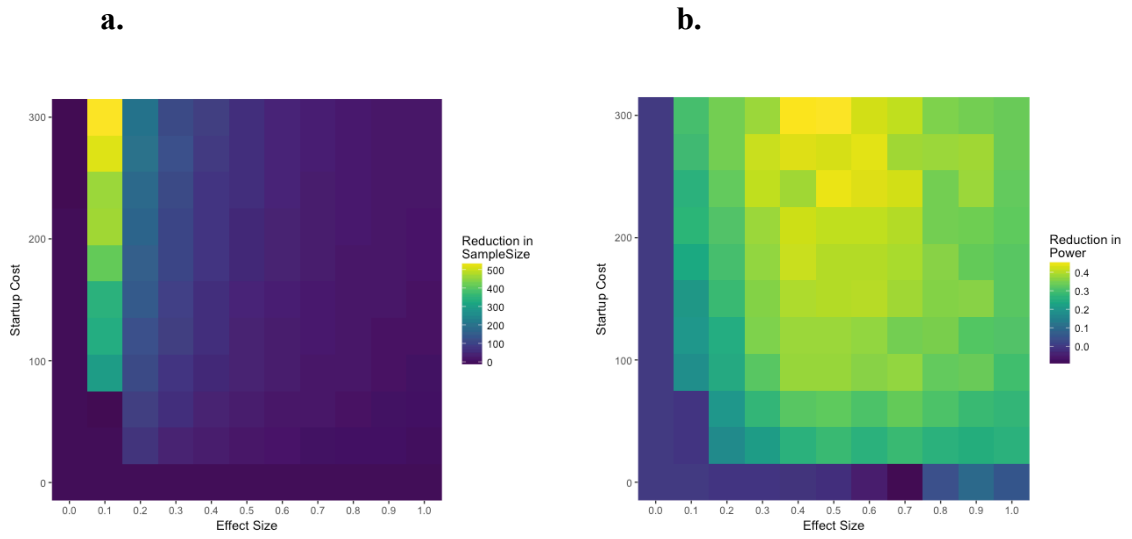


Figure 4S | Reduction in (a) sample size and (b) statistical power due to competition relative to individual optimum (50 generations); 2 competitors. a) When effect sizes are large and startup costs are small, competition has a smaller effect on sample size. The largest reduction in statistical power occurs at high startup costs but intermediate effect sizes. b) Power is most reduced by competition when effect sizes are intermediate, provided that startup costs are non-zero.

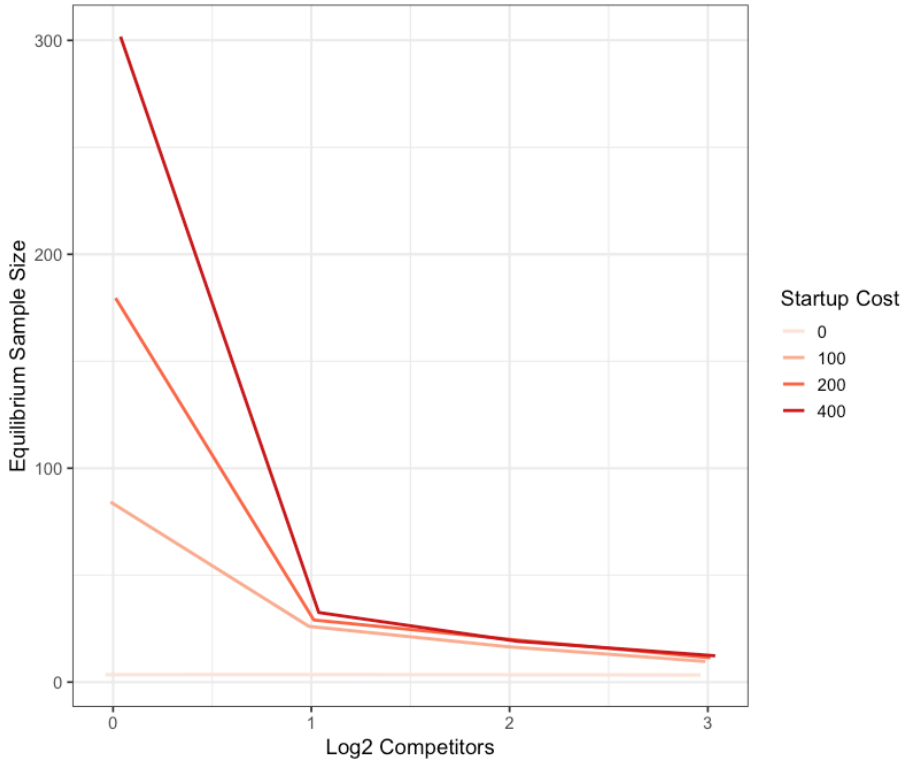


Figure 5S | Alternative plot of results in Figure 2: Equilibrium sample size for individual scientists compared to Log2 number of competitors, as a function of startup cost (200 generations, 50 repeats). Parameter values are: n=100, α=0.05, $\lambda = 3$, r=5, T=5000 and c_s=1. For any number of competitors (i.e. 2, 4, 8), equilibrium sample size is lower than that of individual scientists (i.e. competitors = 1). As the number of competitors increases, equilibrium sample size decreases, because more competitors increase the probability that any given researcher will be scooped. As startup costs increase, equilibrium sample size increases. The effect of startup cost on equilibrium sample size is largest when there are few competitors.

APPENDIX D

CO-AUTHOR PERMISSION STATEMENT

The chapters in this dissertation consist of collaborative work in which Leonid Tiokhin is the first listed co-author. All co-authors have granted their permissions for this co-authored work to be used in this dissertation.