Frequency estimation and semantic ambiguity do not eliminate conjunction bias, when it occurs: Replication and extension of Mellers, Hertwig, and Kahneman (2001)

*Subramanya Prasad Chandrashekar
Institute of International Business and Governance,
Lee Shau Kee School of Business and Administration,
Open University of Hong Kong
spchandr@ouhk.edu.hk

*Yat Hin Cheng, *Chi Long Fong, *Ying Chit Leung, *Yui Tung Wong
Department of Psychology, University of Hong Kong, Hong Kong SAR
dixoncyh@connect.hku.hk; u3510178@connect.hku.hk; u3527150@connect.hku.hk;
crys15@connect.hku.hk

Bo Ley Cheng
Department of Psychology, University of Hong Kong, Hong Kong SAR
boleyc@hku.hk

^Gilad Feldman
Department of Psychology, University of Hong Kong, Hong Kong SAR
gfeldman@hku.hk

In press at Meta Psychology

Accepted for publication on September 10, 2020

*Contributed equally, joint first authors
^Corresponding author: Gilad Feldman
Word: abstract – 150; manuscript - 4537 (excluding tables and figures)

Corresponding author:

Gilad Feldman, Department of Psychology, University of Hong Kong, Hong Kong SAR; gfeldman@hku.hk

Author bios:

Gilad Feldman is an assistant professor with the University of Hong Kong psychology department. His research focuses on judgment and decision-making.

Subramanya Prasad Chandrashekar is a research assistant professor with the Lee Shau Kee School of Business and Administration at the Open University of Hong Kong. His research focuses on social status, lay-beliefs, and judgment and decision-making.

Cheng Yat Hin, Fong Chi Long, Leung Ying Chit, and Wong Yui Tung were students at the University of Hong Kong during the academic year 2018-9.

Bo Ley Cheng was the teaching assistant at the University of Hong Kong psychology department during the academic year 2018-9.

Declaration of Conflict of Interest: The author(s) declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Financial disclosure/funding:

This research was supported by the European Association for Social Psychology seedcorn grant. Subramanya Prasad Chandrashekar would like to thank Institute of International Business and Governance (IIBG), established with the substantial support of a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (UGC/IDS 16/17), for its support.

Authorship declaration

Gilad led the reported replication effort with the team listed below. Gilad supervised each step of the project, conducted the pre-registration, and ran data collection. Prasad followed up on initial work by the other coauthors to verify analyses and conclusions, added equivalence and Bayesian tests with advanced tables and plots, and completed the manuscript submission draft. Prasad and Gilad jointly finalized the manuscript for submission.

Yat Hin Cheng, Chi Long Fong, Ying Chit Leung, and Yui Tung Wong conducted the replication as part of a university course. They conducted an initial analysis of the paper, designed the replication, initiated the extensions, wrote the pre-registration, conducted initial data analysis, and wrote initial replication reports.

Bo Ley Cheng guided and assisted the replication effort.

Contributor Roles Taxonomy

In the table below, employ CRediT (Contributor Roles Taxonomy) to identify the contribution and roles played by the contributors in the current replication effort. Please refer to the url (https://www.casrai.org/credit.html) on details and definitions of each of the roles listed below.

Role	Subramanya Prasad Chandrashekar	Gilad Feldman	Yat Hin Cheng, Chi Long Fong, Ying Chit Leung, and Yui Tung Wong	Bo Ley Cheng
Conceptualization		X		
Pre-registrations		X	X	
Data curation		X		
Formal analysis	X	X	X	
Funding acquisition		X		
Investigation		X	X	
Methodology		X	X	
Pre-registration peer review / verification	X	X	X	X
Data analysis peer review / verification	X		X	
Project administration		X		X
Resources		X		
Software	X	X	X	
Supervision		X		X
Validation	X	X		
Visualization	X			
Writing-original draft	X	X		
Writing-review and editing	X	X		

Abstract

Mellers, Hertwig, and Kahneman (2001) conducted an adversarial collaboration to try and resolve Hertwig's contested view that frequency formats eliminate conjunction effects, and that conjunction effects are largely due to semantic ambiguity. We conducted a pre-registered well-powered very close replication (N = 1032), testing two personality profiles (Linda and James) in a four conditions between-subject design comparing unlikely and likely items to "and" and "and are" conjunctions. Linda profile findings were in support of conjunction effect and consistent with Tversky and Kahneman's (1983) arguments for a representative heuristic. We found no support for semantic ambiguity. Findings for James profile were a likely failed replication, with no conjunction effect. We provided additional tests addressing possible reasons, in line with later literature suggesting conjunction effects may be context-sensitive. We discuss implications for research on conjunction effect, and call for further well-powered pre-registered replications and extensions of classic findings in judgment and decision-making.

Keywords: conjunction effect, frequency estimation, replication, Linda problem, judgment and decision making

Frequency estimation and semantic ambiguity do not eliminate conjunction bias, when it occurs: Replication and extension of Mellers, Hertwig, and Kahneman (2001)

The conjunction fallacy is one of the most well-known judgment errors in the judgment and decision making (JDM) literature. The fallacy consists of judging the conjunction of two events as more likely the any of the two specific events, violating one of the most fundamental tenets of probability theory that postulates that probability of a conjunction of two events can never be higher than the probability any of the two individual events.

Kahneman and colleagues initially reported the conjunction effect as a bias, and that resulted in an intense debate in the academic community (e.g., Fiedler, 1988; Gigerenzer, 1996, 2005; Hertwig & Chase, 1998; Hertwig & Gigerenzer, 1999). One view opposing conjunction effect as a bias was by Hertwig and colleagues that argued that conjunction effect is not at all a fallacy, demonstrating that the effect arises out of semantic ambiguity, in that participants' understanding of natural language words such as "probability" and "and" diverged from that of experimenters (e.g., Hertwig & Gigerenzer, 1999). Daniel Kahneman and Ralph Hertwig engaged in an adversarial collaboration to which Barbara Mellers served as an arbiter. They all then jointly examined the potential semantic ambiguity of "and" conjunction to try and explain the conjunction effect reported in the Kahneman and Tversky's study (1996). The article has been influential with over 430 citations according to Google Scholar at the time of writing.

Chosen study for replication: Outline of Mellers et al (2001)

Mellers et al. (2001) conducted examined frequency estimates of personality sketches. They tested two personality sketches in three experiments, one about Linda and the other about James.

For example, the Linda story read as:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.

Participants read the scenario and estimated how many of a 100 people like Linda fit a particular target description. The target descriptions varied between experimental conditions: likely (feminists), unlikely (bank tellers), semantic "and" (bank tellers and feminists), and semantic "and are"' (bank tellers and are feminists). Kahneman argued that the conjunction effect would occur despite frequency estimation was used, reflected from the average frequency estimates of the conjunction conditions "and" and "and are" higher than the unlikely item condition. Hertwig proposed that conjunction phrase "bank teller and are feminists" would not yield support for conjunction effects. The results for the Linda scenario supported Kahneman's prediction across two out of three experiments conducted as part of the adversarial collaboration, whereas, with the James scenario just one experiment supported the prediction.

We summarized findings in the original article in Table 1. The divergence of findings reported across the three experiments made it hard for readers to assess the overall effect size, and we, therefore, conducted a mini meta-analysis summary of their effects across experiments, summarized in Table 2.

Table 1

Summary of findings in Mellers et al. (2001) Experiments 1 to 3 and the replication

		Linda story					**James story**			
	Target	**Exp1**	**Exp2**	**Exp3**	**Replication**	**Target**	**Exp1**	**Exp2**	**Exp3**	**Replication**
Likely target	Feminists	58.1 (2.4)	47.7 (3.4)	47.9 (4.5)	58.43 (1.79)	Artists	41.0 (2.7)	45.1 (2.6)	47.1 (3.3)	36.2 (1.62)
Unlikely target	Bank tellers	24.6 (1.9)	21.4 (2.0)	14.3 (2.9)	9.87 (0.88)	Republicans	28.9 (2.1)	19.8 (1.8)	12.7 (2.6)	18.38 (1.18)
"and"	"and"	**39.9 (2.0)**	**30.4 (2.3)**	**26.4 (3.9)**	**18.8 (1.36)**	"and"	33.1 (1.8)	**42.7 (2.4)**	**22.9 (3.4)**	15.19 (1.15)
"and are"	"and are"	**40.2 (2.7)**	21.8 (2.1)	**22.8 (2.7)**	**19.55 (1.48)**	"and are"	32 (2.5)	20.0 (1.9)	**21.4 (2.7)**	15.55 (1.09)

Table 2

Summary of findings of the original study versus replication

		Original mini-meta	**Replication**		
Scenario	**Comparison**	**Cohen's d with 95% CI**	**T-statistic (one-sided)**	**Cohen's d with 95% CI**	**Replication summary**
Linda Story	"and" and Unlikely target	0.59 [0.36, 0.82]	$t(431.26) = 5.51$, $p < .001$	0.49 [0.31, 0.67]	Signal - consistent
	"and are" and Unlikely target	0.38 [-0.02, 0.77]	$t(419.21) = 5.63$, $p < .001$	0.50 [0.32, 0.67]	Signal - consistent
	"and" and "and are"	0.18 [-0.09, 0.45]	$t(505.55) = -0.37$, $p = .646$	-0.03 [-0.21, 0.14]	No signal-inconsistent (opposite)
James Story	"and" and Unlikely target	0.62 [0.08, 1.15]	$t(507.82) = -1.93$, $p = .973$	-0.17 [-0.35, 0.00]	Signal-inconsistent (opposite)
	"and are" and Unlikely target	0.17 [-0.07, 0.41]	$t(510.69) = -1.76$, $p = .960$	-0.15 [-0.33, 0.02]	No signal-inconsistent (opposite)
	"and" and "and are"	0.41 [-0.26, 1.08]	$t(506.05) = -0.23$, $p = .591$	-0.02 [-0.19, 0.15]	No signal-inconsistent (opposite)

Note. Linda story can be concluded as a successful replication. James replication is a likely failed replication. In addition, there was no support found for semantic ambiguity (comparing "and" and "and are"). In the original article, effect sizes (ES) were not reported; we computed Cohen's d and confidence intervals based on the mean estimates and standard errors of the mean estimates of the outcome variables of the original study (see full tables in supplementary). The effect sizes of the original study presented in the table are based on the mini-meta analysis of Experiment 1, 2, and 3 of Mellers et al. (2001), as the study is closest for direct comparison for replication summary. The replication summary directly based on LeBel et al., (2019) category, see details in "evaluation criteria for replication design and findings".

The need for replication

Since the first demonstration of the conjunction effect, there have been attempts to develop a theory to explain the phenomenon. Semantic ambiguity remains the strongest counterargument to the demonstration of conjunction effects. With the recent growing recognition of the importance of reproducibility and replicability in psychological science (e.g., Brandt et al., 2014; Open Science collaboration, 2015; van't Veer & Giner-Sorolla, 2016; Zwaan, Etz, Lucas, & Donnellan, 2018), we felt it was important to establish the replicability of the findings noted in the Mellers et al. (2001).

We, therefore, embarked on a well-powered preregistered very close replication of Mellers et al. (2001) employing the most current psychological science methods, which would allow to test for both the presence and possible absence of an effect.

Present investigation

We had several goals. First, we set out to revisit the original experimental design and assess the replicability of the original findings. With power analyses and higher power, we aimed at detecting weak effects that may not have been possible in the original study. Secondly, we complemented the traditional analyses in the original article with equivalence tests and Bayesian analyses to also allow for quantifying evidence in support of the null hypothesis. Third, we added extensions to examine further lay perceptions of provided statistical information that may explain some of the differences found in the original findings.

Context: Large replication effort of judgement and decision-making findings

The current replication was part of a large-scale pre-registered replication project aiming to revisit well-known research findings in the area of judgment and decision making (JDM) and to examine the reproducibility and replicability of these findings. In this project, all replications are conducted by students in undergraduate courses and undergraduate and masters guided thesis at the University of Hong Kong psychology department. Four students in two separate courses were randomly assigned to the current replication. Working independently, the students conducted an in-depth analysis of the target article, wrote pre-registrations with power-analyses, conducted data analysis on the collected data, and then wrote manuscripts for journal submission. In each student pair, students conducted peer review on one another to optimize design and analysis. A teaching assistant (6th author) and the corresponding author supervised and gave feedback in each step of the replication process. The corresponding author conducted all pre-registrations on the OSF and online data collection. More information on the process is provided in the supplementary, and further details and updates on this project can be found on: https://osf.io/5z4a8/ (CORE, 2020).

Method

Pre-registration, power analysis, and open-science

We pre-registered the experiment on the Open Science Framework (OSF), and data collection was launched later that week. Pre-registration with power analyses and all materials used in the study are available in the supplementary materials. All measures, manipulations, and exclusions are reported, and data collection was completed before analyses. OSF pre-registration review link for the study: https://osf.io/gb7pk . Data and R/RMarkdown code (R Core Team,

2015) is available on the OSF: https://osf.io/6v8e2/ . Full open-science details and disclosures are provided in the supplementary.

We aimed to detect smallest the effect size of $d = 0.20$ at a power of 0.80 one-tail comparing two conditions, despite the reported effects in the target article and original findings being much higher. This was meant to allow us the possibility of detecting effects not found in the target article for one of the two scenarios (details below).

Participants

A total of 1032 participants were recruited online through American Amazon Mechanical Turk (MTurk) using the TurkPrime.com platform (Litman, Robinson, & Abberbock, 2017) (M_{age} = 38.77, SD_{age} = 12.07; 550 females). We identified four responses to be excluded based on the exclusion criteria we recorded in the pre-registration due to their self-reported lack of seriousness or English proficiency, yet exclusions had no impact on the findings and so our main report focuses on the full sample.

Procedure

Participants were randomly assigned to one of the four experimental conditions (likely, unlikely, "and", and "and are"). All participants read two personality profiles, one of Linda and the other of James, exactly as in the original study. Each profile consisted of one short description of a character, and frequency estimation questions.

All descriptions and questions were taken from the original article (Mellers et al., 2001). The presentation order of the two profiles was randomized. Linda profile description was as follows:

> Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.
>
> Of 100 people like Linda, how many are [likely: feminists?] [unlikely: bank tellers?] ["and": bank tellers and feminists?] [“and are”: bank tellers and are feminists?]

James profile description was as follows:

> James grew up in a Bohemian family. His father was a musician, and his mother was a painter. They lived together for 40 years and never got married. James was a very talented child with a special gift for comedy, but he turned into a rebellious troublemaker in his youth. He dropped out of college after two years and traveled to Asia to learn crafts. James is now 35 years old.
>
> Of 100 people like James, how many are [likely: artists?] [unlikely: Republicans?] [“and”: Republicans and artists?] [“and are” Republicans and are artists?]

Participants answered questions based on two scenarios, one for Linda and one for James, according to their randomly assigned condition (indicated in brackets in the scenarios above). The dependent variable was the estimated frequency of the described personality in the scenario measured on a scale from 1 to 100. The supplementary details the experimental instructions, scenarios, and response variables.

Extension

Following the replication materials, participants proceeded to the next page and answered six additional questions. Depending on their assigned condition participants were asked to estimate the percentage of people, females, and males in the United States that match the target

item (likely, unlikely, "and", "and are"), and they did so for both profiles. For example, participants in the likely condition estimated the percentage of people, females, and males in the United States that are 1) feminists, 2) artists.

We had several aims with this extension: 1) assess whether the conjunction effect would show for the generalized population without the specific descriptions of James and Linda, and 2) examine possible gender differences in the estimations of the items used in the James and Linda descriptions.

Data analysis plan

Our analyses matched the original article's hypotheses, as follows:

> *Hypothesis 1:* The frequency estimate for the “and” conjunction phrase will be higher than the phrase describing unlikely target alone.

Two sets of competing hypotheses suggested by Hertwig and Kahneman:

> *Hypothesis 2a:* The frequency estimate for the “and are” conjunction phrase will be higher than the phrase describing unlikely target alone.
>
> *Hypothesis 2b:* The frequency estimate for the “and are” conjunction phrase will not be higher than the phrase describing unlikely target alone.
>
> *Hypothesis 3a:* The frequency estimate for the “and are” conjunction phrase will be lower than the frequency estimate for ‘and” conjunction phrase.
>
> *Hypothesis 3b:* The frequency estimate for the “and are” conjunction phrase will not be lower than the frequency estimate for ‘and” conjunction phrase.

A comparison of the three experiments in the original article and the current replication is provided in Table S4 of the Supplementary Materials. In Table S5, we briefly note the reasons for the chosen differences between original studies and the replication attempt. In the replication attempt, we did not include filler items, because when filler items are present, the responses are inherently comparative and therefore drive the conjunction effect observed (Hertwig & Chase, 1998). Supporting this view, the results of both Study 1 and Study 3 of the original study that included filler items found support for conjunction effect—for both "and" and "and are" conjunction phrases. Given the possibility of different psychological processes between comparative and non-comparative responses, we excluded filler items, that allow for the test of competing predictions from Kahneman and Hertwig theorized to be essentially non-comparative in nature. More importantly, with the current focus on testing the main argument if the conjunction effects are driven by semantic ambiguity of natural language term "and" in a frequency representation.

We chose to focus on "and" and "and are" as the conjunction phrases and implement a between-subjects design which would allow for a clearer test of the competing predictions between Kahneman and Hertwig. For instance, Hertwig argued that the frequency judgments are possibly driven by the understanding that "and" is a union operator, and the use of a more restrictive "and are" phrase would take away the conjunction effect. Kahneman argued that judgments were driven by a match between a personality description and porotype of a category; therefore, both "and" and "and are" phrases would likely yield conjunction effects.

Following the analyses in the target original, we first conducted Welch (based on recommendations of Delacre, Lakens, & Leys, 2017) one-tail independent samples t-test, a null-hypothesis significance testing (NHST) method. When NHST analyses were non-significant, we

complement NHST analyses with equivalence testing to compare effects against a minimal effects considered meaningful (TOSTER package; Lakens, 2017; Lakens, Scheel, & Isager, 2018) and Bayesian analyses to quantify support for the null hypothesis given a prior (Kruschke & Liddell, 2018; Vandekerckhove, Rouder, & Kruschke, 2018) using BayesFactor R package (Version 0.9.12-4.2; Morey & Rouder, 2015). These were minor adjustments we made to the pre-registration data analysis plan, summarized in Table S6.

Evaluation criteria for replication design and findings

Table S7 provides a classification of the replications using the criteria by LeBel, McCarthy, Earp, Elson, and Vanpaemel (2018) criteria (see Figure S2). We summarize the current replication as a "very close replication".

To interpret the replication results we followed the framework by LeBel, Vanpaemel, Cheung, and Campbell (2019). They suggested a replication evaluation using three factors: (a) whether a signal was detected (i.e., confidence interval for the replication Effect size (ES) excludes zero), (b) consistency of the replication ES with the original study's ES, and (c) precision of the replication's ES estimate (see Figure S1).

Results

Descriptive statistics are detailed in Table 1 and statistical tests and effect-size findings are summarized in Table 2.

Conjunction effects

We first looked for the conjunction effect for each profile, by comparing frequency estimates for both "and" and "and are" conditions with the "unlikely" condition. Considering the Linda scenario, "and" condition (n = 252, M = 18.80, SD = 21.62) were greater than for the

"unlikely" condition (n = 258, M = 9.87, SD = 14.1; M_d = 8.93, t(431.26) = 5.51, p < .001, d_s = 0.49, 95% CI [0.31, 0.67]; see Figure 1). Similarly, frequency estimates of "and are" condition were greater than "unlikely" condition (n = 258, M = 19.87, SD = 14.15; M_d = 9.69, t(419.21) = 5.63, p < .001, d_s = 0.50, 95% CI [0.32, 0.67]). Thus, results lend support toward H1 and H2a in the Linda scenario.

However, differences across conditions for the James scenario (see summary plot in Figure 1; "and" condition: n = 252, M = 15.19, SD = 18.24; "unlikely" condition: n = 258, M = 18.38, SD = 19.03; "and are" condition: n = 258, M = 15.55, SD = 17.55). The "and" versus "unlikely" contrast (M_d = −3.19, t (507.82) = −1.93, p = .973; d_s = -0.17, 95% CI [-0.35, 0.00]) show that frequency estimates for "and" condition were lower than "unlikely" condition, although the difference was not statistically significant. Therefore, the results of the James scenario failed to support H1. Similarly, the contrast between "unlikely" and "and are" conditions (M_d = −2.83, t(510.69) = −1.76, p = .960; d_s = -0.15, 95% CI [-0.33, 0.02]) show that frequency estimates for "and are" condition were lower than "unlikely" condition, though with a weak effect not statistically significant. In essence, the results support H2b.

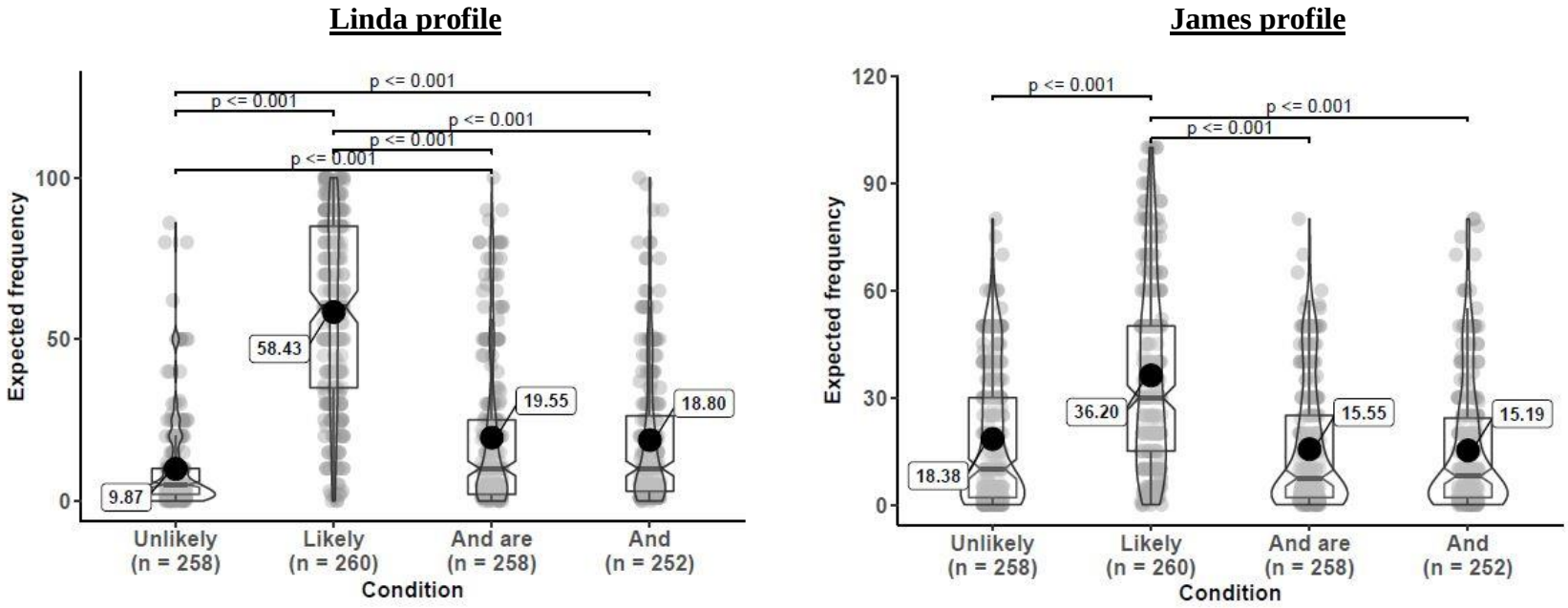


Figure 1. **Linda and James profiles: violin plots for expected frequency of target item.**
Boxes represent interquartile range of the distribution, with the notch in the middle representing the mean. The density of the violin plots represents the density of the data at each value, with wider sections indicating higher density. Note that the p-values for the contrast effects are for two-tail tests, different from the one-tail tests. Plots were generated using ggstatsplot R package (Patil, 2018).

Semantic ambiguity?

To examine whether the semantically ambiguous word "and" had an effect on participants' judgment, we conducted a one-tail Welch t-test comparing frequency estimates of "and" and "and are" conditions for each of the personality scenarios. As predicted by H3a, we found no support for differences for the Linda profile (M_d = −0.75, t(505.55) = −0.37, p = .646, d_s = -0.03, 95% CI [-0.21, 0.14]) or for the James profile (M_d = -0.36, t(506.05) = -0.23, p = .591, d_s = -0.02, 95% CI [-0.19, 0.15]).

Next, we conducted an equivalence test of the semantic ambiguity effect. Based on Simonsohn's (2015) recommendation for replication studies we calculated the smallest effect size of interest (SESOI) that Mellers et al.'s experiment could have detected with a power of 33%. We choose Experiment 2 of as a reference for equivalence test analysis based on one important similarity between the Experiment 2 and the current replication. That is, both studies did not include filler items. With an N of 96 in each condition, Mellers et al. (2001) had 33% power to detect an effect size of d = 0.22. We used it as the equivalence bound for the Study (SESOI set to d = 0.22). Equivalence tests for both Linda story (t(505.55) = -2.21, p = .018) and James story (t(506.05) = -2.25, p = .012) indicating support for the null, meaningfully smaller from SESOI.

Furthermore, we conducted one-tail Bayesian t-tests with a prior set at 0.707 with a null region of (0, ∞) such that the results against null (i.e., against mu = 0) would quantify support the semantic ambiguity hypothesis suggested by Hertwig and colleagues. For the Linda profile, we found BF_{10} = 0.08 (or BF_{01} = 13.32), which indicates that, given the data, the null-hypothesis is over 11 times more likely than the one-sided alternative. Similarly, for the James profile, BF_{10} = 0.08 (or BF_{01}= 12.06), which indicates that given data, the null-hypothesis is over nine times more likely than the one-sided alternative.

Additional analyses

The James profile may have been less representative of an artist in comparison to the Linda profile as representative of a feminist. To test this aspect, we compared the average frequency estimations for James and Linda story within 'likely' experimental condition, in which participants rated the extent to which Linda and James were representative of a feminist and an artist, respectively. Frequency estimations for the "likely" condition for Linda profile ("feminists", n = 260, M = 58.43, SD = 28.93) were greater than for James profile ("artists", M = 36.20, SD = 26.08; M_d = 22.22, t (259) = 11.99, p < .001, d_s = 0.74, 95% CI [0.61, 0.88]). Whereas, a similar comparison between Linda and James story within the unlikely condition show that frequency estimate for Linda ("Bank teller", n = 258, M = 9.87, SD = 14.15) was lower than James ("Republicans", M = 18.38, SD = 19.03; M_d = −8.52, t (257) = −6.87, p < .001, d = -0.43, 95% CI [-0.56, -0.30]). This pattern of the observed difference between Linda and James across "likely" and "unlikely" conditions is consistent with the previous work that found that the occurrence of conjunction effects, for example, depends on the probabilities of A (*Linda is a bank teller)* and B (*Linda is active in the feminist movement*). In particular, there is a higher chance of conjunction effect when people perceive lower the probability of the less probable constituent P(A), and P(B) was high, in comparison to cases where P(A) and P(B) were both low or both high (Fisk & Pidgeon, 1996; Wells, 1985).

The study included additional variables that mirrored the outcome variables but asked the participants to rate the percentage of males and females in the population that fit the description. For example, participants in 'and' condition after reading Linda story answered "Try and estimate, what percentage of females in the U.S. are Bank Tellers and Feminists?", and after reading James story answered "Try and estimate, what percentage of males in the U.S. are

Republicans and Artists?". We looked at the contrasts between the outcome variables and these additional variables across experimental conditions to ascertain if the ratings on the outcome variable were driven by profile description, rather than Linda by virtue of the name being female and similarly James being male. For Linda story across three experimental conditions Linda was rated higher on the outcome variable in comparison to the percentage of females in society (likely condition: M_d = 15.31; t (259) = 8.67, p < .001; d = 0.54, 95% CI [0.41, 0.67]; 'and' condition: M_d = 6.43; t (251) = 4.75, p < .001; d = 0.30, 95% CI [0.17, 0.43]; 'and are' condition: M_d = 5.79; t (257) = 3.98, p < .001; d = 0.25, CI [0.12, 0.37]). Similarly, for the James story, across conditions we found that James was rated higher on the outcome variable in comparison to the percentage of males in society (likely condition: M_d = 19.10; t (259) = 11.15, p < .001; d = 0.69, CI [0.56, 0.83]; 'and' condition: M_d = 3.81; t (251) = 3.36, p = .001; d = 0.21, 95% CI [0.09, 0.34]; 'and are' condition: M_d = 2.58; t (257) = 2.39, p = .018; d = 0.15, 95% CI [0.03, 0.27]).

Summary of replication findings

The evaluation of the replication findings is summarized in Table 2. Our replication for the Linda profile was in support of the confirmatory predictions based on the conjunction effects. Whereas, the results for the James profile were inconsistent. Importantly, the original study reported that in frequency estimate for "and" condition is higher than Unlikely condition. This prediction forms the basis for testing the absence or presence of semantic ambiguity in predicting the conjunction effects. The replication results for this prediction are in the opposite direction, i.e., we found frequency estimates were lower for Unlikely condition than "and" condition. Therefore, the results of the James scenario are inconclusive in teasing apart the semantic ambiguity associated with "and" conjunction term.

Extension

Descriptive results for the extension are provided in Table S8, and plots are provided in Figures S3 to S6.

We first tested whether the conjunction effect occurred for any of the three items (people, male, females; within design) for each of the profiles (Linda and James, between design) and their assigned condition (likely, unlikely, "and", "and are"). As expected, we found no support for a conjunction effect for general population females with the Linda profile items (feminist and bank teller) yet without the Linda description. Similarly, we found no effect for males with the general population James profile items (Republicans and artist) yet without the James description. These findings should be interpreted with caution, yet these are in support of the conjunction effect demonstrated with the Linda and James problems as being affected by the description of Linda and James in a way that makes conjunction items more salient than the unlikely. Meaning, that the conjunction effect may be dependent on the representativeness heuristic (Tversky & Kahneman, 1982) and the preceding described profile.

Yet, we found support for a conjunction effect for the Linda items for the estimation of people overall (feminist: M = 29.36, SD = 17.13; bank teller: M = 8.56, SD = 12.2; "and": M = 11.01, SD = 14.01). It remains to be explored why there would be support for a conjunction effect for evaluation of people overall, but not for females or males, yet it does point out that the conjunction effect may sometimes occur without the representativeness heuristic description, and with a within-subject design. At the very least, this suggests that the conjunction effect is context-sensitive, as is also indicated in the differences in effects we found between the Linda and the James problem.

There were also patterns indicating statistical flaws, such that given a population gender split of 50%-50% for females-males, participants indicated means for the general population that were far from the average of the estimation for females and the estimation of males (e.g., people who are bank teller: $M = 8.56$, $SD = 12.2$; females who are bank tellers: $M = 21.46$, $SD = 28.64$; males who are bank tellers: $M = 9.93$, $SD = 15.40$). This is despite the within-subject design and the three questions being presented together. If participants indeed understood these questions correctly, this may be indicative of elicitation of estimate separately for each of the questions irrespective of the context or priors, and/or an inability to process or report percentages.

Further findings regarding gender effects for the items in the two profile is provided in Tables S10 and S11.

Discussion

We conducted a preregistered well-powered replication of the main design across the three studies of Mellers et al.'s (2001).

Our findings regarding the Linda profile demonstrate support for conjunction effects for both "and" and "and are" connectors. The findings of the Linda scenario are not supportive of the alternative view that that conjunction effects observed in the Linda story are a manifestation of semantic interpretation of "and" term by participants as union instead of the intersection. The semantic ambiguity arguments predicted that "and are" experimental condition will fail to provide support for conjunction effects, and participants' frequency estimate in "and are" experimental condition will be lower than "and" experimental condition. Furthermore, in reference to Linda story, we compared if the frequency estimates in the "and are" condition was lower than "and" condition. Equivalence testing and Bayesian analyses indicated support for null

differences. These findings are in support of the Kahneman view of conjunction effects with frequency estimates.

Our findings for the James profile were not in support of either the Kahneman or the Hertwig hypotheses and previous findings. Firstly, the comparison between “and” and “unlikely” condition was not in support of a conjunction effect. Secondly, we found no support for differences between frequency estimates between “and are” an unlikely condition. Further, similar to Linda story the planned comparison that tested if the frequency estimates in the “and are” condition was lower than “and” condition supports the view that differences between conditions were statistically equivalent to zero. Failure to find empirical support for conjunction effects with James story suggests that conjunction effects are context specific. Conjunction effects are commonly demonstrated using the Linda profile, yet the findings regarding other scenarios are less clear (Costello & Watts, 2017). Thus, it is quite possible that James and Linda scenarios are qualitatively different.

A closer examination of the original findings showed that the effects of the James scenario varied considerably across the experiments from weak effects in Experiment 1 ("and" and unlikely: $d = 0.21$; "and are" and unlikely: $d = 0.13$) with no indication of semantic ambiguity ($d = 0.05$) to mixed effects in Experiment 2 ("and" and unlikely: $d = 1.11$; "and are" and unlikely: $d = 0.01$) indicating strong semantic ambiguity effect ($d = 1.08$). The mini meta-analytic effect we computed for the three original studies seemed to indicate differences in effect size between the Linda and the James scenarios, especially in regards to semantic ambiguity.

Additional analyses we conducted suggested that the personality sketch of James was less representative of an artist in comparison to Linda’s personality sketch of a feminist. The observed difference is consistent with view Kahneman’s argument that conjunction effects arises through

the substitution of representativeness estimates for probability estimates. This may have been one of the reasons why the current study does not find support for conjunction effect for James story even when then comparison was between the unlikely and the "and" conditions, which was supported in Study 2 and 3 of the original paper.

The current replication effort supports the Tversky and Kahneman's (1983) assertion that conjunction effects, when those occur, are a probabilistic error due to representativeness and availability heuristic. More precisely, the results of the current study for Linda story are supportive of the view that frequency estimates do produce conjunction effects that rely on judgmental heuristic and are not driven by semantic ambiguity of the conjunction terms. The results for the James profile were inconclusive to likely failure.

Overall, we found some support for conjunction effects, but that those may be less robust than initially expected. These findings indicate the importance of further conducting well-powered pre-registered replications and extensions that would revisit classic experiments in this domain and aim to gain deeper insights of effect, to investigate the reliability and generalizability of previous findings, the contextual variations of the conjunction effect.

References

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. DOI: 10.1016/j.jesp.2013.10.005

Delacre, M., Lakens, D., & Leys, C. (2017). Why Psychologists Should by Default Use Welch's *t*-test Instead of Student's *t*-test. *International Review of Social Psychology*, *30*, 92–101. DOI: http://doi.org/10.5334/irsp.82

Collaborative Open-science REsearch (2020). Large-scale replications and extensions of findings in Judgment and Decision Making. DOI 10.17605/OSF.IO/5Z4A8. Retrieved March 2020 from http://osf.io/5z4a8

Costello, F., & Watts, P. (2017). Explaining high conjunction fallacy rates: The probability theory plus noise account. *Journal of Behavioral Decision Making*, *30*, 304-321. DOI: 10.1002/bdm.1936

Fiedler, K. (1988). The dependence of the conjunction fallacy on subtle linguistic factors. Psychological Research, 50, 123–129. DOI: 10.1007/BF00309212

Fisk, J. E., & Pidgeon, N. (1996). Component probabilities and the conjunction fallacy: Resolving signed summation and the low component model in a contingent approach. *Acta Psychologica*, *94*, 1-20. DOI: 10.1016/0001-6918(95)00048-8

Gigerenzer, G. (1996). On narrow norms and vague heuristics: A reply to Kahneman and Tversky (1996). Psychological Review, 103, 592–596. DOI: 10.1037/0033-295X.103.3.592

Gigerenzer, G. (2005). I think, therefore I err. *Social Research: An International Quarterly*, *72*, 195-218.

Hertwig, R., & Chase, V. M. (1998). Many reasons or just one: How response mode affects reasoning in the conjunction problem. Thinking and Reasoning, 4, 319–352. DOI: 10.1080/135467898394102

Hertwig, R., & Gigerenzer, G. (1999). The 'conjunction fallacy' revisited:How intelligent inferences look like reasoning errors. *Journal of Behavioral Decision Making,* 12, 275–305. DOI: 10.1002/(SICI)1099-0771(1999)

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, *25*, 178-206. DOI: 10.3758/s13423-016-1221-4

Lakens, D. (2017). Equivalence tests: a practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, *8*, 355-362. DOI: 10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259-269. DOI: 10.1177/2515245918770963

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science, 1*(3), 389-402. DOI: 10.1177/2515245918787489

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). *A Brief Guide to Evaluate Replications.* Meta Psychology, 541, 1–17. DOI: 10.31219/osf.io/paxyn

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*, 433-442. DOI: 10.3758/s13428-016-0727-z

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*, 269-275.DOI: 10.1111/1467-9280.00350

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: Computation of Bayes factors for common designs (R Package Version 0.9.12-2). Retrieved from https://CRAN.R-project.org/package=BayesFactor

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, *349*, aac4716–aac4716. DOI: 10.1126/science.aac4716

Patil, I. (2018). ggstatsplot:"ggplot2" Based Plots with Statistical Details. CRAN.

R Core Team (2015) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07- 0, URL http://www.R-project.org

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*, 559-569. DOI: 10.1177/0956797614567341

Tversky, A., & Kahneman, D. (1982). Judgments of and by representativeness. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. UK Cambridge: Cambridge University Press.

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review*, *90*, 293-315. DOI: 10.1037/0033-295X.90.4.293

Vandekerckhove, J., Rouder, J. N., & Kruschke, J. K. (2018). Bayesian methods for advancing psychological science. *25*, 1-4. DOI: 10.3758/s13423-018-1443-8

van‘t Veer, A.E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, 67, 2-12. DOI: 10.1016/j.jesp.2016.03.004

Wells, G. L. (1985). The conjunction error and the representativeness heuristic. *Social Cognition, 3*, 266-279. DOI: 10.1521/soco.1985.3.3.266

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

Mellers et al. (2001) conjunction effect replication & extension: Supplementary

Contents

Open Science

Data and code

Data and code are shared using the Open Science Framework: https://osf.io/6v8e2/

Pre-registrations and Qualtrics study designs

Link: https://osf.io/gb7pk

Procedure and data disclosures

Data collection

Data collection was completed before analyzing the data.

Conditions reporting

All collected conditions are reported.

Data exclusions

Details are reported in the materials section of this document

Variables reporting

All variables collected for this study are reported and included in the provided data.

Power analyses

As Mellers, Hertwig & Kahneman's (2001) found no differences few of the contrasts (e.g., "and are" vs. unlikely condition in Experiment 2). We aimed to detect a weak effect size of d= 0.2 or above at 8 % power. Therefore, using G*Power alpha = .05, one-tail (direction of hypothesis known), d = 0.2 and power .80 we required a sample size of 310 participants in each condition, and a total sample size of 1240 participants. The final sample was 1028. Please refer to details of the power analysis described below.

Details of the power analysis

Data of the original article

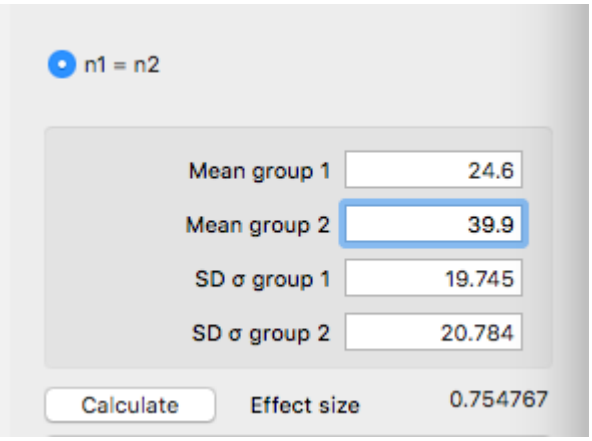
- Power analysis of the current replication will be based on relevant conditions, which are "Likely Target", "Unlikely Target", "and" and "and are" conditions.
- Average frequency estimates for the experiments are shown below, with standard errors and calculated standard deviations.
- Independent T-tests are used to compare the conjunction conditions with the Unlikely Target. Boldface indicates significant results, $p<0.05$
- Numeric result of the T-test is not reported

	Linda's Problem	James' Problem
Likely Target	$\bar{x}$=58.1 SE=2.4 SD = 24.94	$\bar{x}$=41.0 SE=2.7 SD = 28.06
Unlikely Target	$\bar{x}$=24.6 SE=1.9 SD = 19.75	$\bar{x}$=28.9 SE=2.1 SD = 21.82
"and"	**$\bar{x}$=39.9** **SE=2.0** **SD = 20.78**	$\bar{x}$=33.1 SE=1.8 SD=18.71
"and are"	**$\bar{x}$=40.2** **SE=2.7** **SD = 28.06**	$\bar{x}$=32.0 SE=2.5 SD = 25.98

(Table 2 from Millers et al., 2001, p.272)

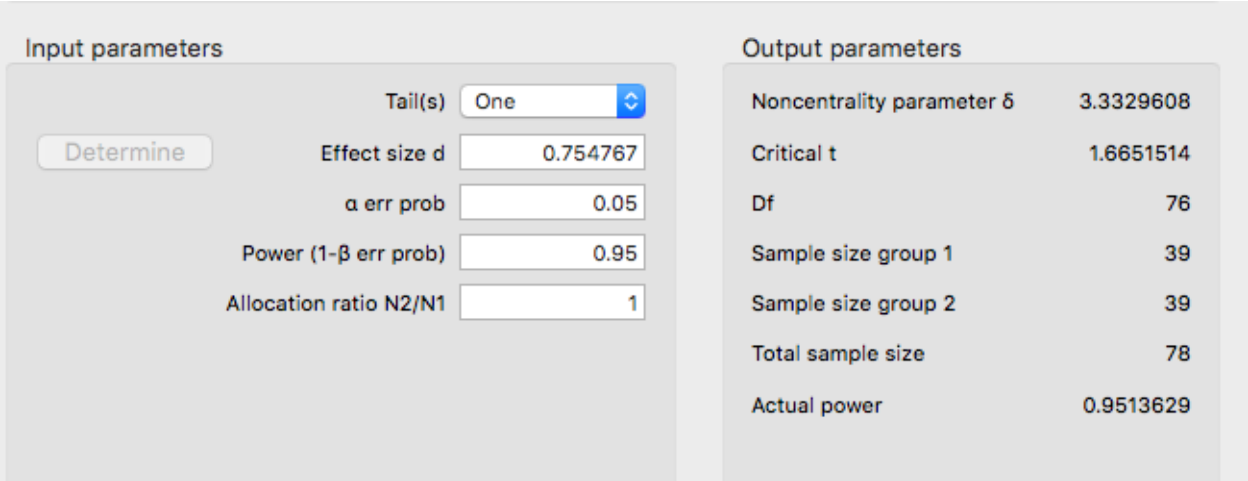
Power Analysis and Sample Size Calculation

Linda's Problem: Unlikely items condition; "And" Condition



Sample Size Calculation

The calculated effect size is plugged in. The calculated sample size will be based on a power of 0.95.



t tests - Means: Difference between two independent means (two groups)

Analysis: A priori: Compute required sample size

Input: Tail(s) = One

Effect size d = 0.754767

α err prob = 0.05

Power (1-β err prob) = 0.95

Allocation ratio N2/N1 = 1

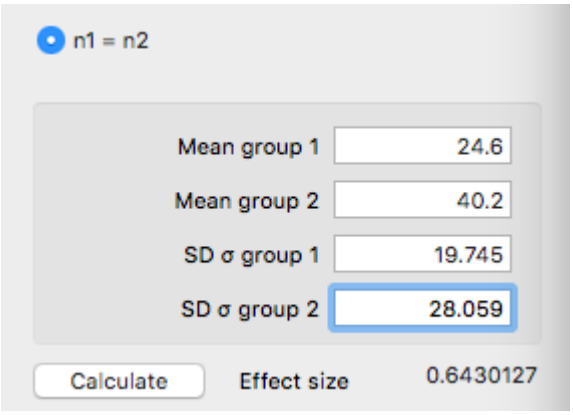
Output: Noncentrality parameter δ = 3.3329608

Critical t = 1.6651514

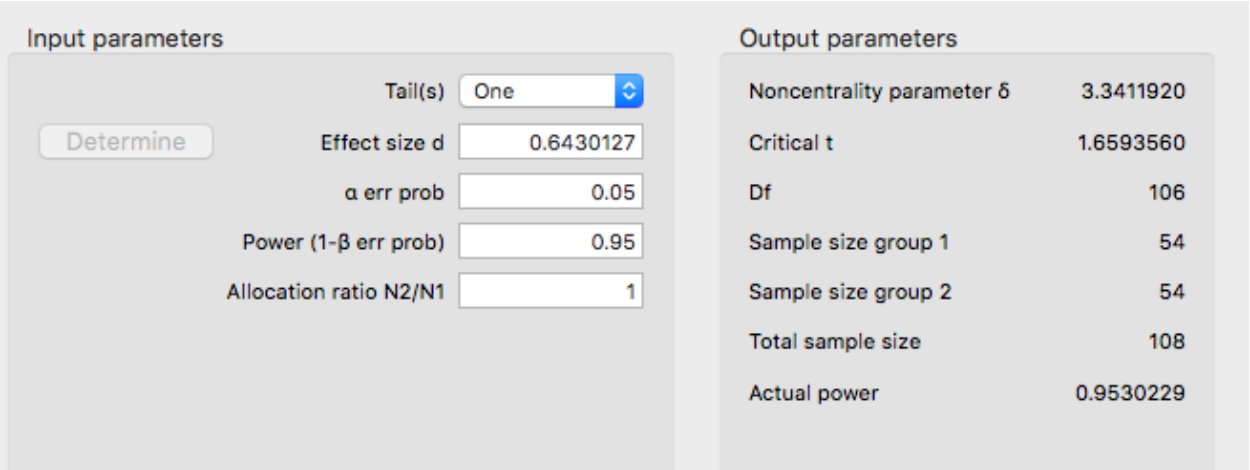
Df	=	76
Sample size group 1	=	39
Sample size group 2	=	39
Total sample size	=	78
Actual power	=	0.9513629

Linda's Problem: Unlikely items; "And are" Condition

Effect Size calculation



Sample size calculation:



t tests - Means: Difference between two independent means (two groups)

Analysis: A priori: Compute required sample size

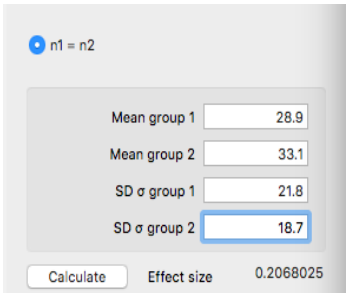
Input:

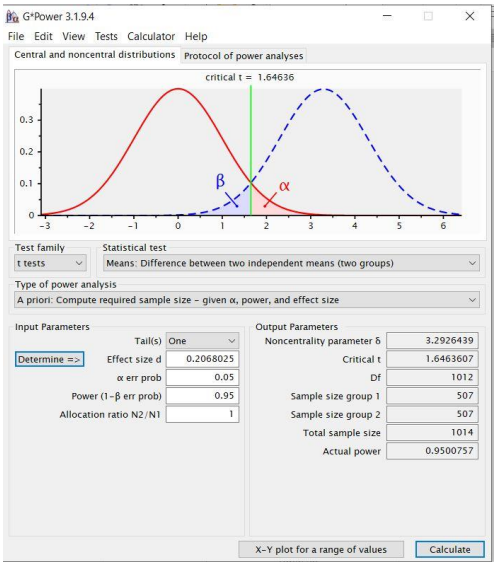
Tail(s)	=	One
Effect size d	=	0.6430127
α err prob	=	0.05
Power (1-β err prob)	=	0.95

	Allocation ratio N2/N1	=	1
Output:	Noncentrality parameter δ	=	3.3411920
	Critical t	=	1.6593560
	Df	=	106
	Sample size group 1	=	54
	Sample size group 2	=	54
	Total sample size	=	108
	Actual power	=	0.9530229

James' Problem: Unlikely items condition; "And" Condition

Effect Size Calculation



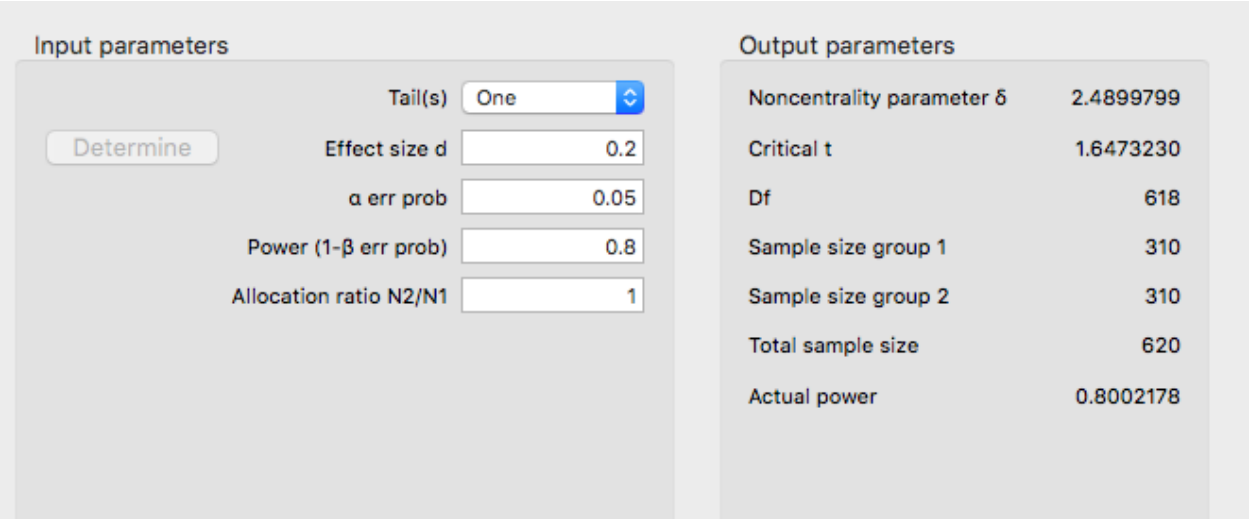


James' Problem

Effect Size Calculation

- No conjunction effect was found in James scenario in the original study
- The current replication hypothesized a very small effect size, Cohen's *d* = 0.2, between "unlikely item" condition and "and" condition, as well as between "unlikely item" condition and "and are" condition for sample size calculation.

Sample Size Calculation



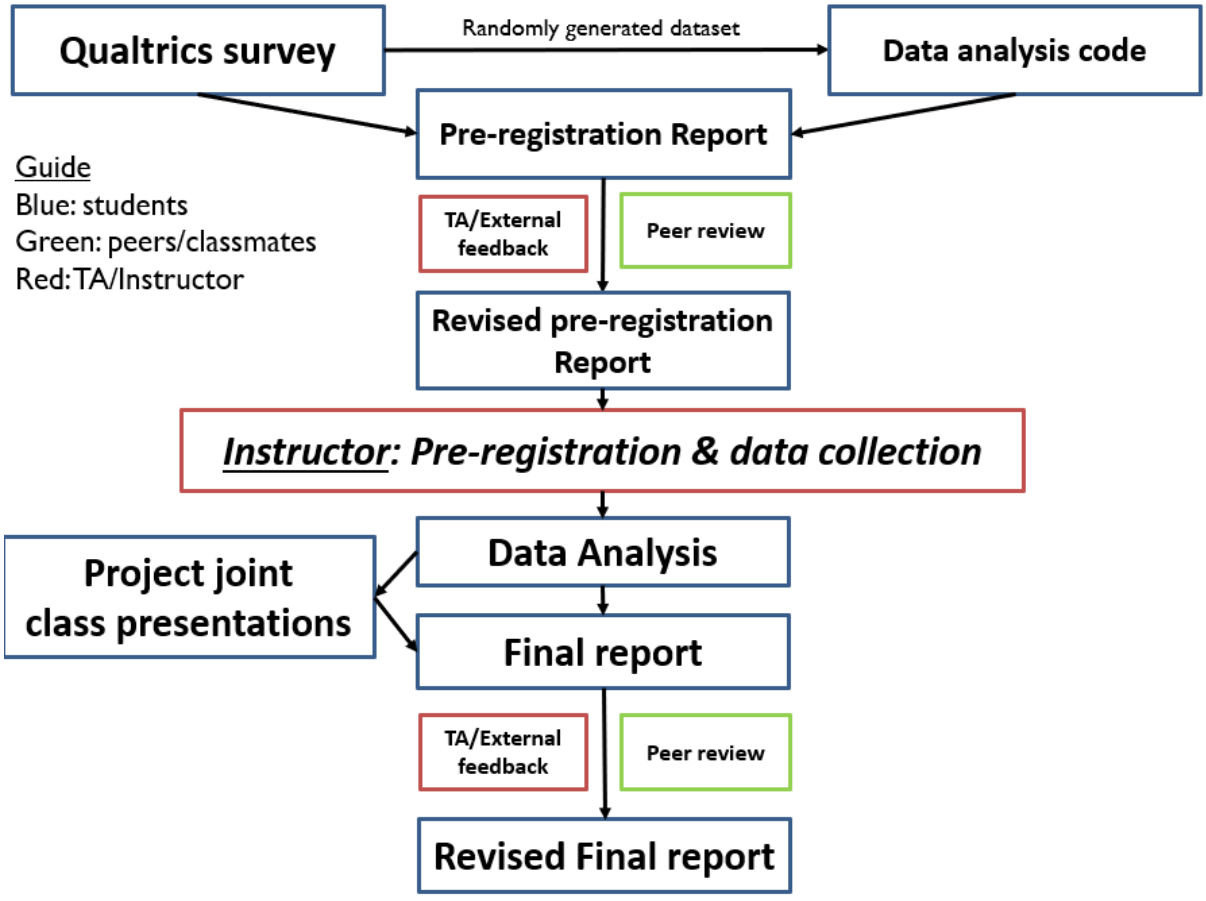
To ensure the effect size of the original study is captured, a more conservative sample size which require at least 310 participants in each condition is considered. It makes a total of 1240 participants.

Project Process Outline

The current replication is part of the mass pre-registered replication project, with the aim of revisiting well-known research findings in the area of judgment and decision making (JDM) and examining the reproducibility and replicability of these findings.

The current replication followed the same project outline as noted below. For each of the replication projects, researchers completed full pre-registrations, data analysis, and APA style submission ready reports. Each of these four researchers (second to fifth author) independently reproduced the materials and designed the replication experiment, with a separate pre-registration document. The researchers then peer-reviewed one another to try and arrive at the best possible design. Then, then last two authors reviewed the integrated work and the last corresponding author made final adjustments and conducted the pre-registration and data collection.

The OSF page of the project contains one Qualtrics survey design used for data collection with four pre-registration documents submitted by each of the researchers. In the manuscript, we followed the most conservative of the four pre-registrations.



Verification of Analyses

Initial analyses were conducted by the independent researchers, who were used JAMOVI (JAMOVI project, 2018) in the analyses. In preparing this manuscript, the lead and corresponding authors verified the analyses in R. T-tests were conducted using base R package, point estimates and confidence intervals for Cohen's d were calculated using 'esc' or 'effsize' R package. We aggregated the effect sizes across the studies using 'meta' R package.

Additional Tables and Figures

Moved from main manuscript to keep manuscript short and concise.

Tables

Table S1

Summary of findings in Mellers et al. (2001) Experiments 1 to 3 and the replication

		Linda story					**James story**			
	Target	**Exp1**	**Exp2**	**Exp3**	**Replication**	**Target**	**Exp1**	**Exp2**	**Exp3**	**Replication**
Likely target	Feminists	58.1 (2.4)	47.7 (3.4)	47.9 (4.5)	58.43 (1.79)	Artists	41.0 (2.7)	45.1 (2.6)	47.1 (3.3)	36.2 (1.62)
Unlikely target	Bank tellers	24.6 (1.9)	21.4 (2.0)	14.3 (2.9)	9.87 (0.88)	Republicans	28.9 (2.1)	19.8 (1.8)	12.7 (2.6)	18.38 (1.18)
“and”	“and”	**39.9 (2.0)**	**30.4 (2.3)**	**26.4 (3.9)**	**18.8 (1.36)**	“and”	33.1 (1.8)	**42.7 (2.4)**	**22.9 (3.4)**	15.19 (1.15)
“and are”	“and are”	**40.2 (2.7)**	21.8 (2.1)	**22.8 (2.7)**	**19.55 (1.48)**	“and are”	32 (2.5)	20.0 (1.9)	**21.4 (2.7)**	15.55 (1.09)
[Unlikely target] "who are" [likely target]	Bank tellers who are feminists	**34.6 (2.3)**	23.1 (2.2)			Republicans who are artists	26.5 (2.2)	24.0 (2.6)		
[Likely target] "who are" [unlikely target]	Feminists who are bank tellers	27.6 (2.2)				Artists who are Republicans	29.5 (2.5)			
Targets combined with no conjunction	Feminist bank tellers	**32.3 (2.3)**				Republican artists	26.0 (2.2)			

Note. Standard errors are in parentheses. Boldface indicates significant results comparing to the unlikely target at $p < .05$.

Table S2

Summary of calculated effect-sizes in Mellers et al. (2001) Experiments 1, 2, and 3

		Effect size (Cohen's d with 95% CI (two-sided))			
Scenario	**Comparison**	**Experiment 1**	**Experiment 2**	**Experiment 3**	**Mini meta-analytic effect size**
Linda Story	"and" and Unlikely target	0.76 [0.48, 1.03]	0.43 [0.14, 0.71]	0.56 [0.12, 1.01]	0.59 [0.36, 0.82]
	"and are" and Unlikely target	0.65 [0.37, 0.92]	0.02 [-0.26, 0.30]	0.49 [0.04, 0.93]	0.38 [-0.02, 0.77]
	"and" and "and are"	-0.01 [-0.28, 0.25]	0.40 [0.11, 0.69]	0.17 [-0.27, 0.61]	0.18 [-0.09, 0.45]
	Bank tellers who are feminists' and Unlikely target	0.46 [0.19, 0.73]	0.08 [-0.20, 0.37]		
	Feminists who are bank tellers' and Unlikely target	0.14 [-0.13, 0.41]			
	Feminist bank tellers' and Unlikely target	0.35 [0.08, 0.62]			
James Story	"and" and Unlikely target	0.21 [-0.06, 0.48]	1.11 [0.80, 1.41]	0.54 [0.09, 0.99]	0.62 [0.08, 1.15]
	"and are" and Unlikely target	0.13 [-0.14, 0.40]	0.01 [-0.27, 0.29]	0.53 [0.08, 0.97]	0.17 [-0.07, 0.41]
	"and" and "and are"	0.05 [-0.22, 0.32]	1.08 [0.77, 1.38]	0.08 [-0.36, 0.52]	0.41 [-0.26, 1.08]
	Republicans who are artists' and Unlikely target	-0.11 [-0.37, 0.16]	0.19 [-0.09, 0.48]		
	Artists who are republicans' and Unlikely target	-0.06 [-0.33, 0.21]			
	Republican artists' and Unlikely target	0.15 [-0.12, 0.42]			

Note. Mellers et al. (2001) included 108, 96, and 40 participants per condition in Experiment 1, Experiment 2, and Experiment 3, respectively. The original study did not report the effect sizes (ES). ES was calculated based on the mean and standard error of the outcome variables across between-subject conditions.

Table S3

Descriptive statistics of findings of the replication

		Linda story					James story				
Experimental condition	***n***	**Target item or conjunction phrase**	***Mean***	***SD***	***Skewness***	***Kurtosis***	**Target item or conjunction phrase**	***Mean***	***SD***	***Skewness***	***Kurtosis***
Likely	260	Feminists	58.43	28.93	-0.28	-1.06	Artists	36.20	26.08	0.58	-0.59
Unlikely target	258	Bank teller	9.87	14.15	2.71	8.62	Republicans	18.38	19.03	0.89	-0.33
"and"	252	"and"	18.80	21.62	1.54	1.91	"and"	15.19	18.24	1.52	1.74
"and are"	258	"and are"	19.55	23.74	1.48	1.19	"and are"	15.55	17.55	1.21	0.49

Table S4

Comparison between original and replication study

	Original Study			**Current Replication Study**
	Experiment 1	**Experiment 2**	**Experiment 3**	
Number of Conditions	Seven conditions, five of them are conjunctions that involve both "likely" and "unlikely" items.	Five conditions, which include three conditions that employed of conjunction phrases.	Four conditions, which include two conditions that employed of conjunction phrases.	Four conditions, which include two conditions that employed of conjunction phrases.
Filler items	One filler item Included	Not included	Five filler item Included	Not included
Design	Between-subjects	Between-subjects	Between-subjects	Between-subjects
Number of experimental conditions	7	5	4	4
Sample size	756 (Average of 108 per experimental condition)	480 (Average of 96 per experimental condition)	160 (Average of 40 per experimental condition)	1028 (Average of 257 per experimental condition)
Participants population	Undergraduates at Ohio State University.	Undergraduates at Ohio State University.	Undergraduates at University of California, Berkeley	Participants from Amazon Mechanical Turk (MTurk).
Remuneration	Course credits	Monetary reward	Not reported	Monetary reward

Table S5

Difference and similarities between original studies and the replication attempt

	Original Study	**Replication Study**	**Reason of changes**
Number of Conditions	Four to seven conditions, five of them are conjunctions that involve both "likely" and "unlikely" items, see Tables 1 and 2.	Four conditions, two involve the use of conjunction phrases, which are "and' and "and are".	We choose the four between experimental conditions that are common across all three studies of Mellers et al. (2001). The chosen experimental conditions test the most important arguments that surround the conjunction effect.
Filler items	Included in Study 1 and Study 3	Not included	The findings from the original study suggested that the filler items contributed toward the differences in the results between Linda and James story. Thus, we avoided the filler items. (For the same reasons the Study 2 of the original work did not include filler items in Study 2)
Procedure	Not reported	Online survey using Qualtrics surveying platform.	Allows minimal error in data collection and entry, and useful in faster data collection.
Participants population	Undergraduates at the University.	Online Amazon Mechanical Turk of varied demographic background.	To recruit more participants, more diverse population of a wider demographic range.
Sample Size	Average of 108, 96, and 40 participants per condition in Study 1, 2, and 3, respectively.	1028 across four experimental conditions (257 participants/condition)	See power analysis in supplementary.

Table S6

Preregistration planning and deviation documentation

Components of pre-registration	Were there deviations?	If yes describe the details of the deviation(s)	Rationale for deviation	How might the results be different if had not deviated
Procedures	No	N/A	N/A	N/A
Power analysis	No	N/A	N/A	N/A
Exclusion rules	No	N/A	N/A	N/A
Evaluation criteria	Minor additions	For evaluation of the replication we employed the LeBel et al.'s (2018) framework.	The framework aids researchers to conduct systematic evaluation of credibility of empirical findings.	N/A
Analysis	Minor additions	We conducted Equivalence tests and Bayesian analysis were performed in addition to null-hypothesis significance tests (NHST).	Both Equivalence test and Bayesian analysis are useful in testing for and quantifying an absence of an effect	With the additional tests along with NHST tests, we are not only able to falsify predictions about the presence of effects, but also declare the absence of meaningful effects.

Table S7

Classification of the two replication studies based on LeBel et al.'s (2017) taxonomy

Design facet	**Replication study**
IV operationalization	Same
DV operationalization	Same
IV stimuli	Same
DV stimuli	Same
Procedural details	Different (minor adjustments)
Physical settings	Different
Contextual variables	Different
Replication classification	Very close replication

Table S8

Descriptive statistics

Scenario	**Experimental Condition**	***n***	**Median**
Linda	Unlikely	258	5.00
	Likely	260	60.00
	"And are"	258	10.00
	"And"	252	10.00
James	Unlikely	258	10.00
	Likely	260	30.00
	"And are"	258	7.50
	"And"	252	8.00

Table S9

Summary of findings of the original the replication based on 'Mann–Whitney U' test

Scenario	Comparisons	W-statistic; p-value	Effect size (Cliff delta with 95% CI)
Linda	"And" vs. Unlikely	4.13×10^4; $p < .001$	0.27 [0.17, 0.36]
	"And are" vs. "Unlikely"	4.12×10^4; $p < .001$	0.24 [0.14, 0.33]
	"And" vs "And are"	3.32×10^4; $p = .34$	0.02 [-0.08, 0.12]
James	"And" vs. Unlikely	2.94×10^4; $p = .96$	-0.1 [-0.19, 0.01]
	"And are" vs. "Unlikely"	3.08×10^4; $p = 0.93$	-0.07 [-0.17, 0.03]
	"And" vs "And are"	3.19×10^4; $p = .66$	-0.02 [-0.12, 0.08]

Table S10

Summary of findings of the original study 2 versus the replication

		Original Study 2	Replication		
Scenario	**Comparison**	**Cohen's d**	**T-statistic**	**Cohen's d**	**Replication summary**
Linda Story	"and" and Unlikely target	0.43 [0.14, 0.71]	$t(431.26) = 5.51$, $p < .001$	0.49 [0.31, 0.67]	Signal - consistent
	"and are" and Unlikely target	0.02 [-0.26, 0.30]	$t(419.21) = 5.63$, $p < .001$	0.50 [0.32, 0.67]	Signal – inconsistent (larger)
	"and" and "and are"	0.40 [0.11, 0.69]	$t(505.55) = -0.37$, $p = .646$	-0.03 [-0.21, 0.14]	No signal-inconsistent
James Story	"and" and Unlikely target	1.11 [0.80, 1.41]	$t(507.82) = -1.93$, $p = .973$	-0.17 [-0.35, 0.00]	Signal-inconsistent (opposite)
	"and are" and Unlikely target	0.01 [-0.27, 0.29]	$t(510.69) = -1.76$, $p = .960$	-0.15 [-0.33, 0.02]	No signal-consistent
	"and" and "and are"	1.08 [0.77, 1.38]	$t(506.05) = -0.23$, $p = .591$	-0.02 [-0.19, 0.15]	No signal-inconsistent (opposite)

Note. The current table mirrors the comparison noted in the Table 2 of the manuscript, with an important exception. We compare the findings of the replication findings with Study 2 of the original study. In the original article, effect sizes (ES) were not reported; we computed Cohen's d and confidence intervals based on the mean estimates and standard errors of the mean estimates of the outcome variables of the original study (see full tables in supplementary). The replication summary directly based on LeBel et al., (2019) category, see details in "evaluation criteria for replication design and findings".

Figures

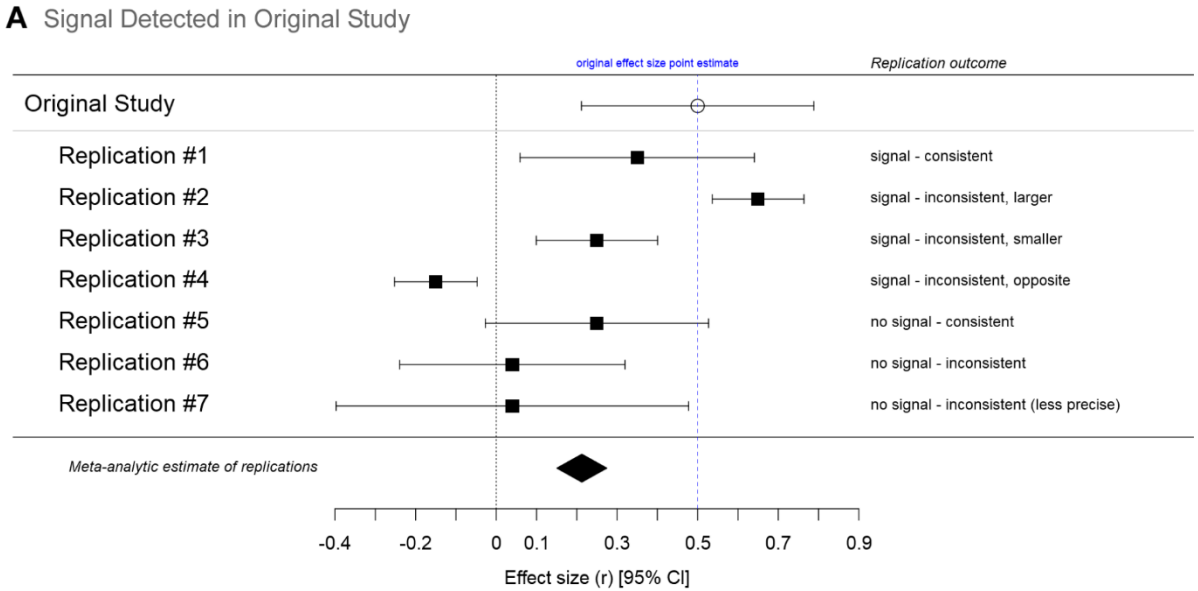


Figure S1. Criteria for evaluation of replications by LeBel et al. (2019). A taxonomy for comparing replication effects to target article original findings.

Target similarity	Highly similar				Highly dissimilar
Category	Direct replication			Conceptual replication	
Design facet	**Exact replication**	**Very close replication**	**Close replication**	**Far replication**	**Very far replication**
IV operationalization	Same	Same	Same	Different	
DV operationalization	Same	Same	Same	Different	
IV stimuli	Same	Same	Different		
DV stimuli	Same	Same	Different		
Procedural details	Same	Different			
Physical setting	Same	Different			
Contextual variables	Different				

Figure S2. Criteria for evaluation of replications by LeBel et al. (2018).
A classification of relative methodological similarity of a replication study to an original study. "Same" ("different") indicates the design facet in question is the same (different) compared to an original study. IV = independent variable. DV = dependent variable. "Everything controllable" indicates design facets over which a researcher has control. Procedural details involve minor experimental adjustments (e.g., task instruction wording, font, font size, etc.).

Materials and scales used in the experiment

Procedure

Participants were randomly assigned to one of the four conditions, and in each condition, read two personality descriptions (Linda and James'). The survey followed the following sequence:

- Participants signed the consent form. Then were given instructions, and then were randomly assign to one of the four conditions, and in each condition read Linda and James Stories one by one. Then rate the frequency for each of the story they read.
- Demographics questions.
- After that, participants filled the funnelling section that checked if they are seriously filling in the survey, and if they can guess the purpose of the study.

Exclusion criteria

In the pre-registration we included the following:

"We will focus on our analyses on the full sample. However, as a supplementary analysis and to examine any potential issues, we will also determine further findings reports with exclusions. In any case, we will report exclusions in detail with results for full sample and results following exclusions (in either the manuscript or the supplementary).

General criteria:

1. Participants indicating a low proficiency of English (self-report<5, on a 1-7 scale)
2. Participants who self-report not being serious about filling in the survey (self-report<4, on a 1-5 scale).
3. Participants who correctly guessed the hypothesis of this study in the funnelling section.
4. Have seen or done the survey before
5. Participants who failed to complete the survey. (duration = 0, leave question blank)
6. Not from United States"

Instructions and experimental material

All participants first read the instruction:

> *We are interested in the judgment and inferences that people make about other peoples' profession, politics, and hobbies. In each of the following problems, we will tell you about a person. We will then ask, of 100 people like the target person, how many would fit a particular description of a job, political persuasion, or hobby?*
>
> *There will be two scenario and questions. Please read the introduction and the items carefully. There are no right or wrong answers; please answer to the best of your understanding.*

After that, participants were randomly assigned to one of four experimental conditions and in each condition read Linda Story and James story. The order of Linda and James story was randomized. Linda story read:

Please read the following description and answer the questions stated below:

> *Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice, and also participated in anti-nuclear demonstrations.*

Participants then proceed to read the James Story:

> *James grew up in a Bohemian family. His father was a musician, and his mother was a painter. They lived together for 40 years and never got married. James was a very talented child with a special gift for comedy, but he turned into a rebellious troublemaker in his youth. He dropped out of college after two years and travelled to Asia to learn crafts. James is now 35 years old.*

They then answered the question in a text box (validated to answer in frequency). In each condition, respondents answered two questions, exactly as they did in the original study. After submitting the answers and they proceeded to the next page to answer 6 additional questions.

Experimental condition: likely target

Dependent variables (two questions same as the original study):

- *Of 100 people like Linda, how many are feminist?*
- *Of 100 people like James, how many are Artists?*

Additional questions

- *Try and estimate, what percentage of people in the U.S. are feminists?*
- *Try and estimate, what percentage of females in the U.S. are feminists?*
- *Try and estimate, what percentage of males in the U.S. are feminists?*
- *Try and estimate, what percentage of people in the U.S. are artists?*
- *Try and estimate, what percentage of females in the U.S. are artists?*
- *Try and estimate, what percentage of males in the U.S. are artists?*

Experimental condition: Unlikely target

Dependent variables (two questions same as the original study):

- *Of 100 people like Linda, how many are Bank Tellers?*
- *Of 100 people like James, how many are Republicans?*

Additional questions

- *Try and estimate, what percentage of people in the U.S. are Bank Tellers?*
- *Try and estimate, what percentage of females in the U.S. are Bank Tellers?*
- *Try and estimate, what percentage of males in the U.S. are Bank Tellers?*
- *Try and estimate, what percentage of people in the U.S. are Republicans?*
- *Try and estimate, what percentage of females in the U.S. are Republicans?*
- *Try and estimate, what percentage of males in the U.S. are Republicans?*

Experimental condition: "and"

Dependent variables (two questions same as the original study):

- *Of 100 people like Linda, how many are Bank Tellers and Feminists?*
- *Of 100 people like James, how many are Republicans and Artists?*

Additional questions

- *Try and estimate, what percentage of people in the U.S. are Bank Tellers and Feminists?*
- *Try and estimate, what percentage of females in the U.S. are Bank Tellers and Feminists?*
- *Try and estimate, what percentage of males in the U.S. are Bank Tellers and Feminists?*

- *Try and estimate, what percentage of people in the U.S. are Republicans and Artists?*
- *Try and estimate, what percentage of females in the U.S. are Republicans and Artists?*
- *Try and estimate, what percentage of males in the U.S. are Republicans and Artists?*

Experimental condition: "and are"

Dependent variables (two questions same as the original study):

- *Of 100 people like Linda, how many are Bank Tellers and are Feminists?*
- *Of 100 people like James, how many are Republicans and are Artists?*

Additional questions

- *Try and estimate, what percentage of people in the U.S. are Bank Tellers and are Feminists?*
- *Try and estimate, what percentage of females in the U.S. are Bank Tellers and are Feminists?*
- *Try and estimate, what percentage of males in the U.S. are Bank Tellers and are Feminists?*
- *Try and estimate, what percentage of people in the U.S. are Republicans and are Artists?*
- *Try and estimate, what percentage of females in the U.S. are Republicans and are Artists?*
- *Try and estimate, what percentage of males in the U.S. are Republicans and are Artists?*

Funnelling section

Three funnelling questions:

- *What do you think the purpose of the last part was?*
- *Have you ever seen the materials used in this study or similar before? If yes - please indicate where*
- *Did you spot any errors? Anything missing or wrong? Something we should pay attention to in next runs? (Briefly, up to one sentence, write "none" if not relevant)*

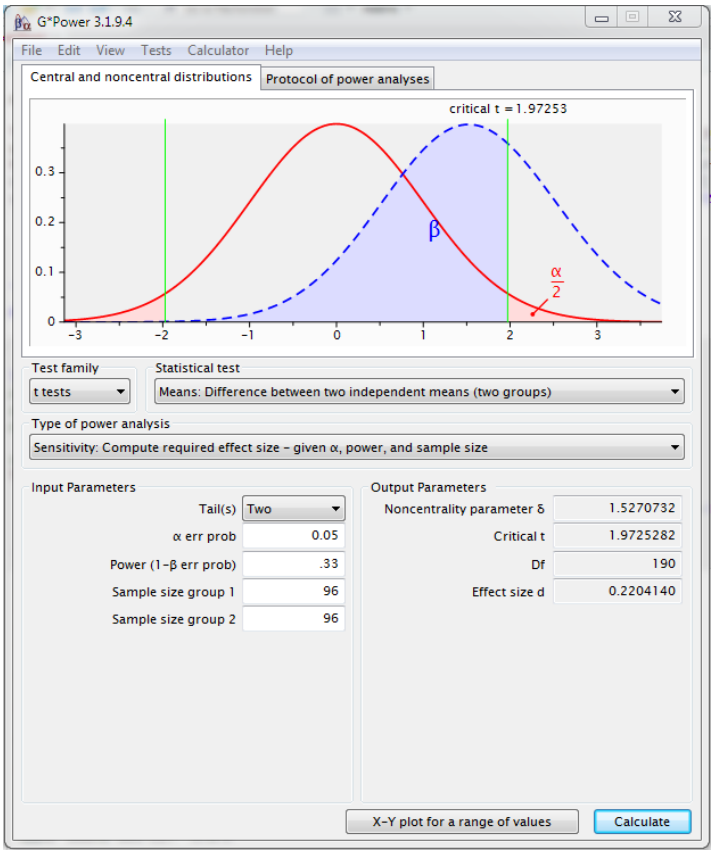
Finally, participants were asked to fill in demographics and were debriefed. No filler items were included.

Additional analyses and results

Calculation of smallest effect size of interest (SESOI) for equivalence test procedure

For calculation of SESOI we choose Experiment 2 of Mellers et al.'s (2001) as the reference as this particular experiment is closest to the current replication design. The average number of participants per condition in Experiment 2 of Mellers et al.'s (2001) study = 96 participants.

With an N of 96 in each condition, Mellers et al. (2001) had 33% power to detect an effect size of 0.22 based on the calculation using the G*power software. The screenshot below demonstrates the details of the calculation:



Estimates of the occurrence of in the populations

To probe into understanding possible reasons for variations in the results between Linda and James story, we conducted exploratory analysis based on the responses to additional variables that were not part of the original study. One way to way to disentangle differences in the results to check for assessment of the lay-perceptions of the occurrences of the prototypes of a category (e.g., feminist bank teller, republican artist) in the population, and to see whether these somehow differ for the two scenarios that may explain the divergence in responding to the two scenarios.

We looked for the possibility of gender as a factor of influence that may have brought about differences in the results between the two scenarios. **Table S3** presents the results of various comparisons that asked participants the percentage of men and women who take up the described roles (e.g., feminists, bank tellers, republicans, artists) in the society. Results of the comparison show that women are significantly more likely to be feminists and bank tellers than men. There was no support for gender difference for the role of artists. However, participants rated that men are significantly more likely to be Republicans than men.

The more appropriate comparison checks if there was a difference in the mean scores of occurrence for “females how are Bank tellers and feminists” and “males who are Republicans and are Artists.” Because, the original study mainly tested the extent to which personality sketches of Linda and James are representative of “females how are Bank tellers and feminists” and “males who are Republicans and Artists,” respectively. The results show that there is no significant difference in the participants’ ratings of general occurrences between “females how are Bank tellers and feminists” and “males who are Republicans and Artists.” To summarize, do not indicate support for the gender effects that possibly explain the difference in results between Linda and James story.

Furthermore, **Table S4** presents the results comparing the population level occurrences of the roles of the personality sketches (e.g., feminists, republicans, artists, bank teller, feminist bank teller, republican artist).

Table S8

Descriptive statistics of the additional measures

Items from	***Variable***	***Mean***	***SD***	***Skewness***	***Kurtosis***
Linda profile	*Try and estimate, what percentage of **people** in the U.S. are feminists?*	*29.36*	*17.13*	*0.45*	*-0.47*
	*Try and estimate, what percentage of **people** in the U.S. are Bank Tellers?*	*8.56*	*12.2*	*3.06*	*14.37*
	*Try and estimate, what percentage of **people** in the U.S. are Bank Tellers and are Feminists?*	***12.29***	***17.04***	*2.39*	*6.33*
	*Try and estimate, what percentage of **people** in the U.S. are Bank Tellers and Feminists?*	***11.01***	***14.01***	*2.07*	*5.01*
Linda profile	*Try and estimate, what percentage of **females** in the U.S. are feminists?*	*43.12*	*23.04*	*0.24*	*-0.86*
	*Try and estimate, what percentage of **females** in the U.S. are Bank Tellers?*	*21.46*	*28.64*	*1.27*	*0.04*
	*Try and estimate, what percentage of **females** in the U.S. are Bank Tellers and are Feminists?*	*13.77*	*18.72*	*1.99*	*3.36*
	*Try and estimate, what percentage of **females** in the U.S. are Bank Tellers and Feminists?*	*12.38*	*17.64*	*2.35*	*5.84*
Linda profile	*Try and estimate, what percentage of **males** in the U.S. are feminists?*	*15.63*	*14.71*	*1.38*	*1.84*
	*Try and estimate, what percentage of **males** in the U.S. are Bank Tellers?*	*9.93*	*15.40*	*2.53*	*8.16*
	*Try and estimate, what percentage of **males** in the U.S. are Bank Tellers and are Feminists?*	*6.32*	*12.49*	*3.30*	*12.26*
	*Try and estimate, what percentage of **males** in the U.S. are Bank Tellers and Feminists?*	*5.17*	*8.94*	*2.80*	*8.15*
James profile	*Try and estimate, what percentage of **people** in the U.S. are artists?*	*20.61*	*16.85*	*1.48*	*2.84*
	*Try and estimate, what percentage of **people** in the U.S. are Republicans?*	*46.01*	*11.51*	*-0.44*	*4.02*
	*Try and estimate, what percentage of **people** in the U.S. are Republicans and are Artists?*	*14.93*	*16.75*	*1.81*	*3.75*
	*Try and estimate, what percentage of **people** in the U.S. are Republicans and Artists?*	*13.86*	*14.46*	*1.34*	*1.2*
James profile	*Try and estimate, what percentage of **females** in the U.S. are artists?*	*18.53*	*17.63*	*1.63*	*2.72*
	*Try and estimate, what percentage of **females** in the U.S. are Republicans?*	*32.39*	*12.71*	*0.78*	*3.05*
	*Try and estimate, what percentage of **females** in the U.S. are Republicans and are Artists?*	*11.33*	*13.83*	*1.88*	*3.34*
	*Try and estimate, what percentage of **females** in the U.S. are Republicans and Artists?*	*9.76*	*11.81*	*1.72*	*2.38*
James profile	*Try and estimate, what percentage of **males** in the U.S. are artists?*	*17.10*	*16.30*	*1.76*	*3.85*
	*Try and estimate, what percentage of **males** in the U.S. are Republicans?*	*46.91*	*16.35*	*-0.19*	*0.01*
	*Try and estimate, what percentage of **males** in the U.S. are Republicans and are Artists?*	*12.98*	*16.65*	*2.17*	*5.28*
	*Try and estimate, what percentage of **males** in the U.S. are Republicans and Artists?*	*11.38*	*14.42*	*1.97*	*3.78*

Table S9

Bayes factor analysis results

Scenario	Comparison	Bayes factor$_{10}$	Bayes factor$_{01}$
Linda Story	“and” and Unlikely target	3.98×10^{5}	2.51×10^{-6}
	“and are” and Unlikely target	6.56×10^{5}	1.52×10^{-6}
	“and” and "and are"	0.08	13.32
James Story	“and” and Unlikely target	0.03	29.07
	“and are” and Unlikely target	0.04	27.29
	“and” and "and are"	0.08	12.06

Table S10

Comparisons based on the additional questions probing for possible gender effects

Comparisons		Type of the test	Statistic	Error ±%	*df*	*p*	Mean difference	SE difference	Cohen's d
Female bankers	Male bankers	Student's t	8.52		249	< .001	11.669	1.37	0.539
		Bayes factor$_{10}$	3.41 x10^{12}	0.00					
Female Republicans	Male Republicans	Student's t	-15.8		249	< .001	-14.91	0.944	-0.999
		Bayes factor$_{10}$	1.22x10^{36}	0.00					
Female feminists	Male feminists	Student's t	23.89		242	< .001	28.065	1.175	1.533
		Bayes factor$_{10}$	1.09x10^{62}	0.00					
Female artists	Male artists	Student's t	2.8		242	0.005	1.601	0.571	0.18
		Bayes factor$_{10}$	3.28	0.00					
Females who are bankers and feminists	Males who are bankers and feminists	Student's t	11.63		982	< .001	3.52	0.303	0.371
		Bayes factor$_{10}$	1.01x10^{26}	0.00					
Females who are republicans and artists	Males who are republicans and artists	Student's t	-3.95		982	< .001	-0.724	0.183	-0.126
		Bayes factor$_{10}$	81.05	0.00					
Females who are bankers and feminists	Males who are republicans and artists	Student's t	1.32		982	0.186	0.453	0.342	0.0422
		Bayes factor$_{10}$	0.09	0.00					

Note: SE = Standard error.

Table S11

Result of t-tests for comparisons at population level

Comparisons		Type of the test	Statistic	Error ±%	*df*	*p*	Mean difference	Cohen's d
People who are bankers	People who are Republicans	Student's t	-40.49		249	< .001	-37.88	-2.561
		Bayes factor$_{10}$	5.0×10^{107}	0.00				
People who are feminists	People who are artists	Student's t	6.62		242	< .001	8.72	0.425
		Bayes factor$_{10}$	3.3×10^{07}	0.00				
People who are bankers and feminists	People who are republicans and artist	Student's t	-4.32		982	< .001	-1.26	-0.138
		Bayes factor$_{10}$	365	0.00				
People who are feminists	People who are Republicans	Student's t	-19.80		429.12	< .001	-16.95	0.425
		Bayes factor$_{10}$	3.3×10^{07}	0.00				
People who are artists	People who are Republicans	Student's t	-12.79		421.22	< .001	-25.67	-1.79
		Bayes factor$_{10}$	2.28×10^{61}	0.00				
People who are artists	People who are bankers	Student's t	9.4		430.89	< .001	12.21	0.85
		Bayes factor$_{10}$	2.91×10^{16}	0.00				

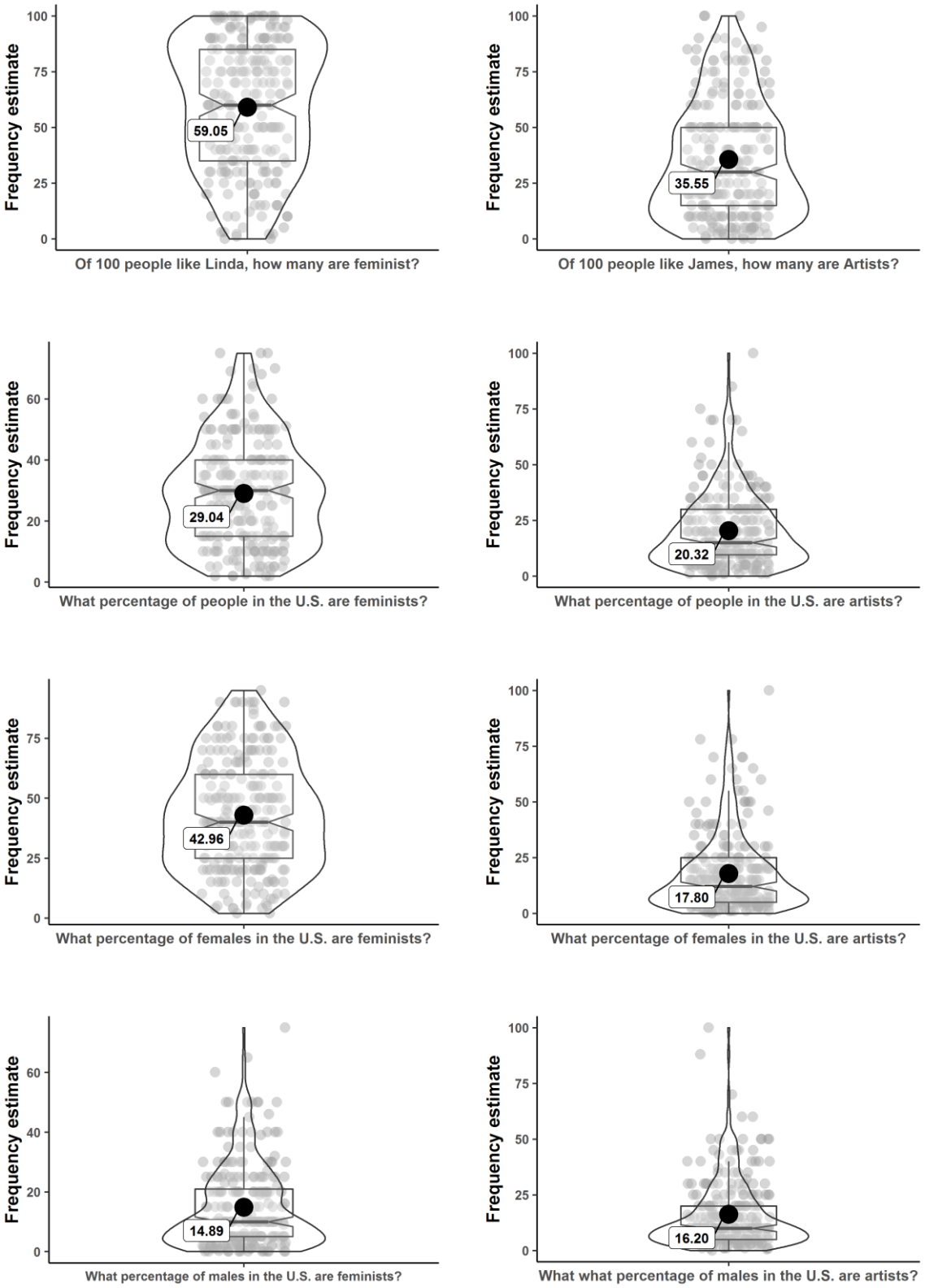


Figure S3. Boxplot and violin-plot with jittered data points of the measures in the "Likely" experimental condition.

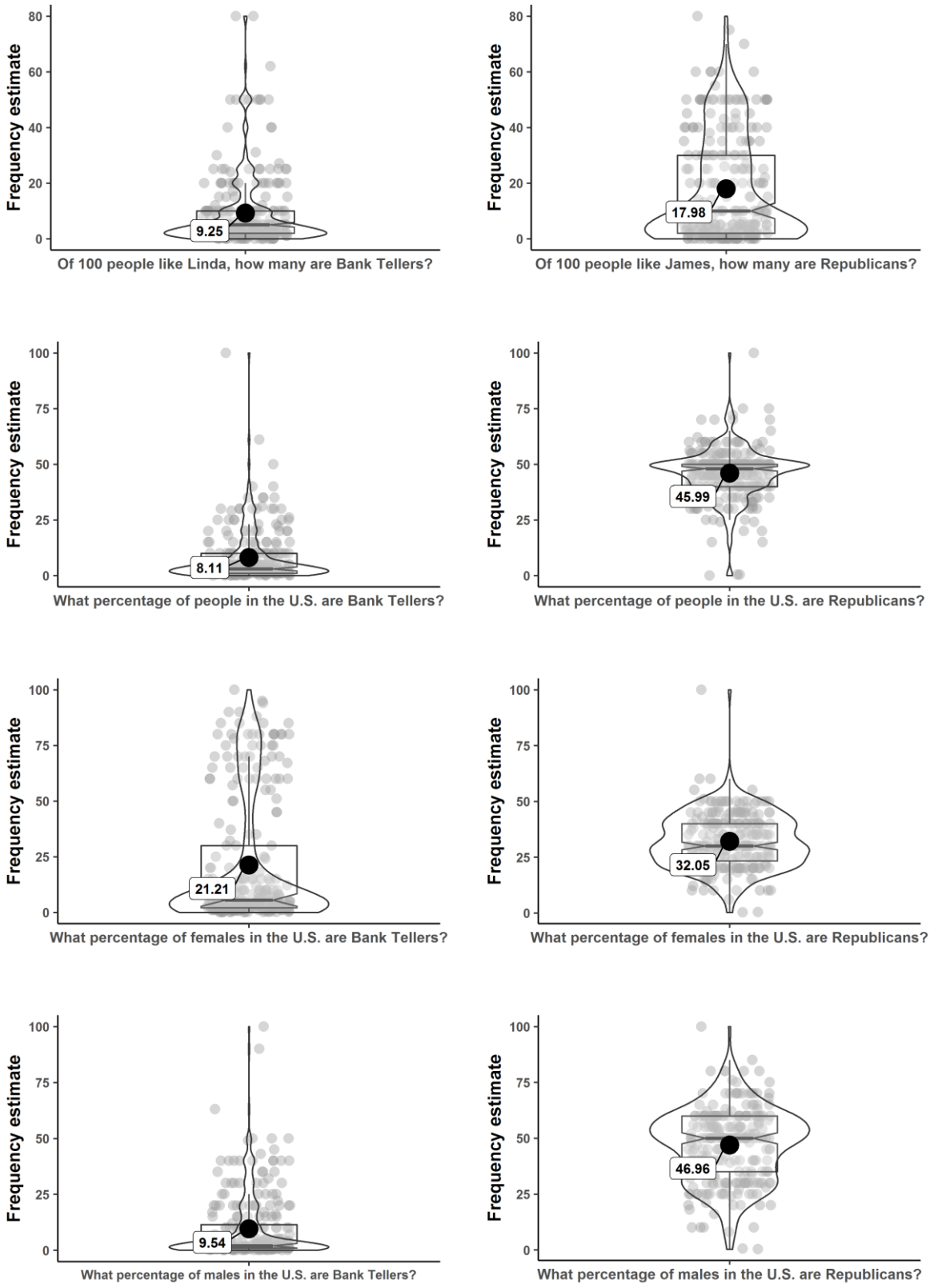


Figure S4. Boxplot and violin-plot with jittered data points of the measures in the "Unlikely" experimental condition.

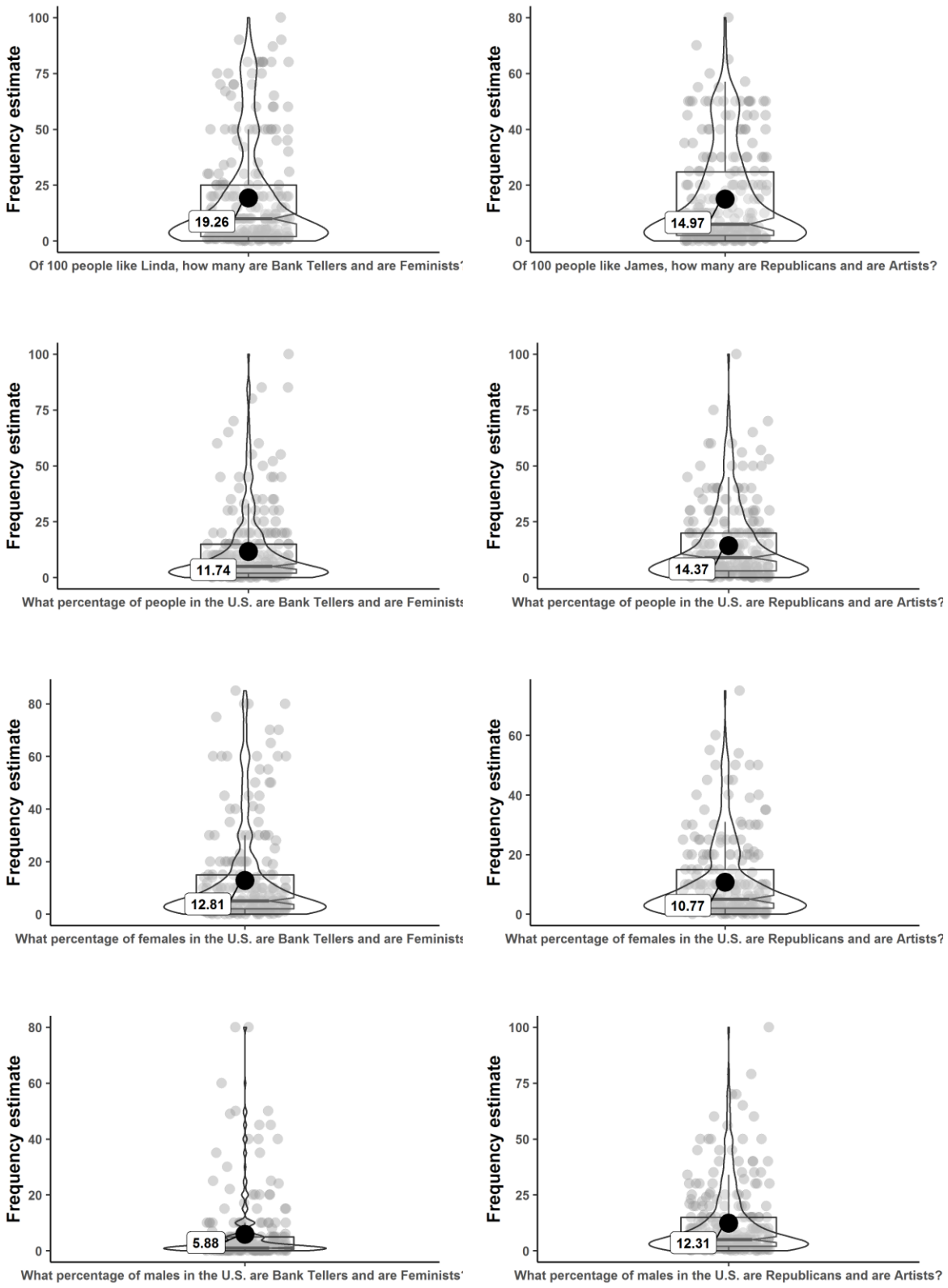


Figure S5. Boxplot and violin-plot with jittered data points of the measures in the "And are" experimental condition.

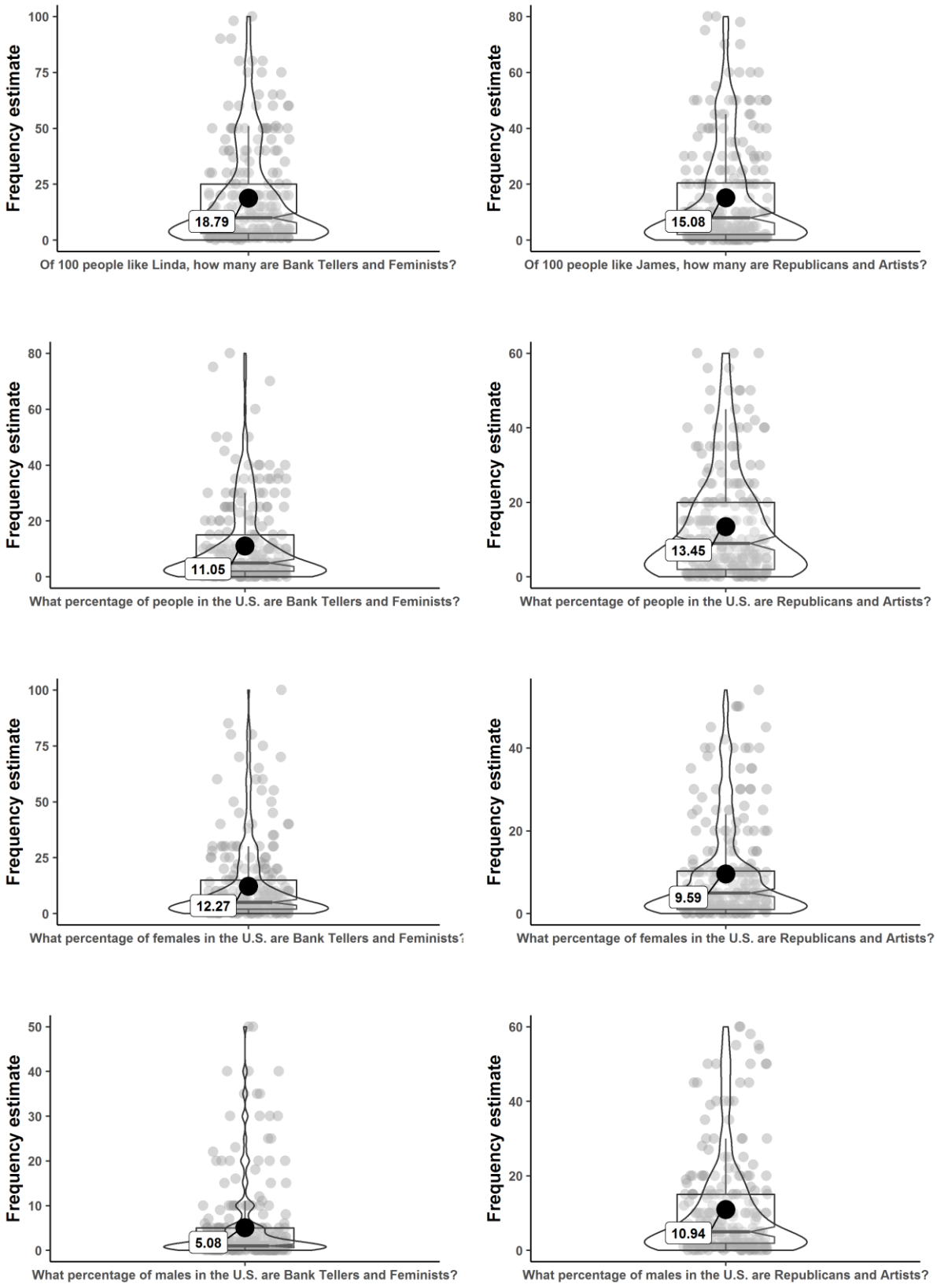


Figure S6. Boxplot and violin-plot with jittered data points of the measures in the "And" experimental condition.

References

Mellers, B., Hertwig, R., & Kahneman, D. (2001). Do frequency representations eliminate conjunction effects? An exercise in adversarial collaboration. *Psychological Science*, *12*(4), 269-275. DOI: 10.1111/1467-9280.00350