**Representativeness heuristic in intuitive predictions:   
Replication Registered Report of problems reviewed in   
Kahneman and Tversky (1973)   
[Stage 1]**

Hong Ching (Bruce) Chan  
ORCID: 0009-0001-8137-8667  
Department of Psychology, The University of Hong Kong  
[u3578465@connect.hku.hk](mailto:u3578465@connect.hku.hk) / [brucechan20020504@gmail.com](mailto:brucechan20020504@gmail.com)

^Gilad Feldman  
ORCID: 0000-0003-2812-6599  
Department of Psychology, The University of Hong Kong  
[gfeldman@hku.hk](mailto:gfeldman@hku.hk) / [giladfel@gmail.com](mailto:giladfel@gmail.com)

^Corresponding author

## 

## Author bios:

Hong Ching (Bruce) Chan was a thesis student at the University of Hong Kong during the academic year 2023-4.

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

## Declaration of Conflict of Interest:

The author(s) declared no potential conflicts of interests with respect to the authorship and/orpublication of this article.

## Financial disclosure/funding:

The project is supported by the University of Hong Kong Teaching Development Grant.

## Authorship declaration:

Hong Ching (Bruce) Chan conducted the replication as part of her thesis in psychology.

Gilad Feldman guided the project, supervised each step in the project, ran data collection, and edited the manuscript for submission.

## Corresponding author

Gilad Feldman, Department of Psychology, University of Hong Kong, Hong Kong SAR; [gfeldman@hku.hk](mailto:gfeldman@hku.hk) ; 0000-0003-2812-6599

## Rights:

CC BY or equivalent license is applied to the AAM arising from this submission. ([clarification](https://bit.ly/rrs-primer))

## Citation of the target research article:

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, *80*(4), 237.

## Acknowledgements (with ORCID)

## Contributor Roles Taxonomy

|  |  |  |
| --- | --- | --- |
| **Role** | **Hong Ching (Bruce) Chan** | **Gilad Feldman** |
| Conceptualization | X | X |
| Pre-registration | X | X |
| Data curation |  | X |
| Formal analysis | X |  |
| Funding acquisition |  | X |
| Investigation | X |  |
| Pre-registration peer review / verification |  | X |
| Data analysis peer review / verification |  | X |
| Methodology | X |  |
| Project administration |  | X |
| Resources |  | X |
| Software | X |  |
| Supervision |  | X |
| Validation |  | X |
| Visualization | X |  |
| Writing-original draft | X |  |
| Writing-review and editing |  | X |

# PCIRR-Study Design Table

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question** | **Hypothesis** | **Sampling plan** | **Analysis plan** | **Statistical tests rationale** | **Interpretation given outcomes** | **Theory affected by the outcomes** |
| Do people make predictions based more on prior probabilities or the representativeness of the available evidence? | People predict based on the representativeness of the evidence to the target of prediction and largely ignore prior probabilities. | We aimed to recruit 800 US Americans through Prolific. Effect sizes and the sample information to calculate effect sizes were not reported in the target article. We aimed for a minimum sample of 800 participants instead (200 participants per experimental condition in Study 4, which has the greatest number of conditions).  Sensitivity analysis indicated that a sample of 800 with a target alpha of .005, we could detect *f* = 0.16 for the 2 by 6 ANOVA in Study 3, *f* = 0.16 for the 2 by 2 ANOVA in Study 4, *d* = 0.39 for the independent samples t-test in Study 5, *d* = 0.16 for the paired samples t-test in Study 6, and *r* = 0.16 for the bivariate correlation in Study 7. | Pearson's correlations for Studies 1 and 2; paired samples t-test, descriptive analysis, and 2 (condition) x 6 (description) ANOVA for Study 3; Levene’s test of homogeneity and 2 (adjectives/ reports) x 2 (evaluation/ prediction) ANOVA for Study 4; independent samples t-tests in Study 5; paired samples t-tests in Study 6; Pearson’s correlation in Study 7. | This is a replication. We follow the methods and analyses of the target article, with additional analyses that seem appropriate for the experimental design. | The replicability of Kahneman and Tversky (1973) was examined based on the replication evaluation criteria by LeBel et al. (2019) | People may not be as neglectful of prior probabilities as Kahneman and Tversky (1973) suggested; the effect of representativeness on the confidence in predictions may not be supported; there might be an association between statistical knowledge and the ability to consider regression to the mean. |
| Does the consistency of available evidence affect the confidence in predictions? | People are more confident in their predictions when the evidence is consistent. |
| Is the ability to consider regression to the mean related to statistical knowledge? | The level of statistical knowledge is not related to the ability to consider regression to the mean. |

# Abstract

[IMPORTANT: Abstract, method, and results were written using a randomized dataset produced by Qualtrics to simulate what these sections will look like after data collection. These will be updated following the data collection. For the purpose of the simulation, we wrote things in past tense, but no pre-registration or data collection took place yet.]

The representativeness heuristic is the phenomenon that people make predictions not by statistically considering prior evidence, but by the representativeness of the evidence to the target of prediction . In a Registered Report experiment with a US American sample on Prolific (*N* = 1309), we conducted a conceptual replication of Studies 1 and 2 and a close replication of Studies 3 to 7 from Kahneman and Tversky (1973). [The following is a demo placeholder based on the random simulated and will be updated following data collection.] We found support for the effects of [...] (effects + 95% CI). We found mixed support for the effects of [...] (effects + 95% CI). However, we failed to find support for the effects of [...] (effects + 95% CI). Extending the replication, we [found/failed to find] support for [...]. Overall, we concluded that [...]. Materials, data, and code are available on: <https://osf.io/8zhcj/>

*Keywords:* representativeness heuristic, bias, judgment and decision making, registered report, replication

# 

# Representativeness heuristic in intuitive predictions: Replication Registered Report of problems reviewed in Kahneman and Tversky (1973) [Stage 1]

[IMPORTANT: Section is written in the past tense to simulate what the manuscript will look like after data collection, yet no pre-registration or data collection took place yet.]

## Background

Kahneman and Tversky (1973) introduced and reviewed the “representativeness heuristic” as a mental shortcut in which people tend to make predictions, evaluations, or classifications more based on representativeness - the degree of resemblance of essential features of the target (e.g., fit with stereotypes, belonging to categories) - than based on objective evidence and statistical information. This is related to yet different from the “availability heuristic”, a mental shortcut in which people tend to make predictions, evaluations, or classifications more based on the ease-of-recall of related examples than based on objective evidence (Tversky & Kahneman, 1973). While both heuristics may be helpful in some circumstances, they may result in systematic biases with real-life implications.

In their review, Kahneman and Tversky (1973) presented the results of seven studies which showed that when making predictions, people shift evaluations towards their predictions of representativeness based on available information of the assessed target. This suggests that people seem to disregard prior probabilities according to Bayes’ theorem and the accuracy or relevance of the evidence. Kahneman and Tversky (1973) concluded that the representativeness heuristic affects both categorical and numerical predictions, is influenced by the consistency of input variables for prediction, and is difficult to avoid despite having relevant statistical knowledge or knowing about the effect. Their seminal work has an immense impact on psychology, economics, policy, and beyond, and is considered the foundation of and related to the Nobel Memorial Prize in Economic Sciences recognition Kahneman received for his work in 2002.

We conducted a replication and extension Registered Report of Kahneman and Tversky (1973) with the following main goals. Our first goal was to conduct an independent close and well-powered replication of the seminal empirical studies initially used to demonstrate the representativeness heuristic. Our second goal was to extend the target article’s design in various aspects, such as in examining people’s confidence in their numerical predictions, and whether statistical knowledge is associated with the persistence of the representativeness heuristic.

We begin by introducing the literature on the representativeness heuristic and the chosen article for replication - Kahneman and Tversky (1973). We then discuss our motivations for the current replication, review Kahneman and Tversky (1973), and outline the chosen studies from the target article, their experimental design, and our adaptations and extensions.

## Representativeness heuristic: Review by Kahneman and Tversky (1973)

Kahneman and Tversky (1973) chose seven studies that demonstrate representativeness heuristic. Their first demonstration (Study 1) pertained to categorical predictions. They first asked a group of participants (the base rates group) to estimate the base rates of enrolment in nine fields of graduate specialization (e.g., business administration, computer science, engineering, etc.). They then proceeded to give two other groups a description of an individual called Tom W. which fits the stereotype of an engineer yet is broad and vague enough to match any profession. Participants in the similarity group ranked the same nine fields in terms of how similar Tom W. is to a typical graduate student in that field, and those in the likelihood group ranked the nine fields in terms of the likelihood that Tom W. is now a graduate student in each of the fields. Their results indicated that the correlation between the likelihood rankings and the similarity rankings was very high and that the correlation between likelihood rankings and the base rates was very low. Kahneman and Tversky (1973) concluded that this meant the participants made predictions almost as if rating how similar (i.e., representative) the description of Tom W. is to the typical student in each field instead of considering the base rates (i.e., the prior probabilities).

Another example of the representativeness heuristic was provided in Kahneman and Tversky’s (1973) third study (Study 3), which investigated the effect of individuating evidence on the neglect of base rates. Imagine a sample of 30 engineers and 70 lawyers, with an individual described as having “high intelligence, although lacking in true creativity. … a need for order and clarity, and for neat and tidy systems in which every detail finds its appropriate place.” (p. 238). When participants were asked about the probability of this person being an engineer from the sample, due to the description resembling that of a typical engineer, participants gave estimates disproportionately higher than 30%. This seems consistent with the idea of the representativeness heuristic. In other words, people would have based their prediction on how the currently available evidence is representative of the target of prediction and disregarded the prior probabilities.

Other studies presented in the target article expanded on the construct of the representativeness heuristic in different aspects: Study 2 also investigated the representativeness heuristic in categorical predictions using the famous Tom W. description, finding that manipulations of expected accuracy had no effect on participants’ predictions; Study 4 (prediction vs evaluation) investigated numerical predictions, finding no support for differences in the variability of participants’ predictions and evaluations, which was consistent with the representativeness hypothesis; Study 5 (prediction vs translation) found support for a potential property of prediction by representativeness wherein participants would translate input values from a scale to the other; Study 6 (effect of consistency) found support for the potential factor of consistency in input evidence which may lead to predictions by representativeness; Study 7 (persistence of non-regressive intuitions) demonstrated how failure to consider regression to the mean may persist despite statistical knowledge. Each of these studies contributed uniquely to the conceptualization of the phenomenon that is the representativeness heuristic. We summarized the setup of Studies 1 to 7, along with the corresponding hypotheses and findings from the target article in Tables 1 and 2.

###### Table 1 *Kahneman and Tversky (1973) Studies 1 to 7: Summary of hypotheses and findings*

| **Study** | **Scenario** | **Hypothesis** | **Findings of the target article** |
| --- | --- | --- | --- |
| 1 | *Similarity and predictions (Tom W.).* Three conditions:  a) Base rates (*n* = 69): Participants asked to consider percentage of first year US students enrolled in nine fields of specialization (e.g., business, computer science, law)  b) Similarity (*n* = 65): Participants were given a personality description of Tom W. and asked to rank the nine fields in order of a typical student’s similarity with Tom W.  c) Likelihood prediction (*n* = 114): Participants were given a personality description of Tom W. and asked to rank the nine fields in order of a typical student’s similarity with Tom W. | The association between the likelihood and similarity rankings is stronger than the association between likelihood and the base rates. | The correlation between likelihood rankings and similarity was .97, whereas that between likelihood and base rates was -.65. |
| 2 | *Similarity and predictions (manipulation of expected accuracy)*. Similar setup as Study 1:  a) Base rate and similarity rating (*n* = 69): Participants were given 5 thumbnail personality sketches of ninth-grade boys allegedly written by a counselor on the basis of an interview in the context of a longitudinal study. Then they were asked to rank how similar it is to their “image of the typical first-year graduate in that field” (p. 240) (the same 9 fields from Study 1 Tom W.) for each of the ninth-grade boys. Afterwards, participants estimated the base rate frequency of 9 fields.  b) Likelihood prediction (high accuracy) (*n* = 55): Participants were told that accuracy in predictions was high and then asked to rank the nine fields according to "the likelihood that the person described is now a graduate student in that field.” (p. 240). Participants also estimated the probability that their first choice was correct.  c) Likelihood prediction (low accuracy) (*n* = 50): Participants were told that accuracy in predictions was low and then asked to rank the nine fields according to "the likelihood that the person described is now a graduate student in that field.” (p. 240). Participants also estimated the probability that their first choice was correct. | a) For each of the five personality sketches, the association between the likelihood and similarity rankings is stronger than the association between likelihood and the base rates.  b) The association between the likelihood rankings and base rates would not be significantly different between the high and low accuracy groups. [null hypothesis]  c) The association between likelihood rankings and base rates is stronger for the null description than the other descriptions. | a) For the law description, the correlation between likelihood rankings and similarity was .93, whilst that between likelihood and base rates was .33; similarly, for the computer science description for that with similarity was .96 and that with base rates was -.35; medicine description: .92 compared to .27; library science description: .88 compared to -.03; business administration description: .88 compared to .62.  b) The correlations between the likelihood rankings and base rates were not significantly different for the high accuracy group (*r* = .13) compared to the low accuracy group (*r* = .16) (*t*(103) = 0.42, *p* = N/A).  c) The correlation between likelihood rankings and base rates for the null description was *r* = .74. |
| 3 | *Prior vs individuating evidence:* Participants were given 5 descriptions (and 1 null description) of individuals supposedly randomly chosen from a sample of 100 engineers/ lawyers.  High engineer condition (*n* = 86): Initial sample was 70 engineers/ 30 lawyers.  Low engineer condition (*n* = 85): Initial sample was 30 engineers/ 70 lawyers.  Participants asked to predict the probability of each description belonging to an engineer. | a) The mean judged probabilities combined from the 5 descriptions are not different between the high and low engineer conditions. [null hypothesis]  b) (Descriptive analysis) On the plot of the median judged probabilities in the high engineer condition against the low engineer condition, only the point representing the null condition (which does not contain any individuating information) would lie on the Bayesian prediction line. | a) The mean judged probabilities combined from the 5 descriptions were different between the high (*M* = 55%) and low (*M* = 50%) engineer condition (*t*(169) = 3.23, *p* < .01). [contrary to expected effect from the hypothesis]  (effect was interpreted as very small through descriptive analyses and taken to show that the median judged probabilities were visually closer to the identity line than the curved line of the correct relation according to Bayes’ rule)  b) (Descriptive analysis) Only the point representing the null condition laid on the Bayesian prediction line. |
| 4 | *Prediction vs evaluation:* 2 (adjectives/ reports) x 2 (evaluation/ prediction) design.  Participants were either given adjectives or reports describing freshmen’s intellectual quality and character. Reports contained additional information about the students’ adjustment to college.  Participants then either asked to evaluate the percentage of students in the entire class whose descriptions indicate a higher academic ability, or to predict the grade point average achieved by each student at the end of his freshman year and his class standing in percentiles. Sample sizes: adjectives-evaluation (*n* = 38), adjectives-prediction (*n* = 36), reports-evaluation (*n* = 37), reports-prediction (*n* = 63). | a) Comparing at the adjectives level, the mean SD is not different between the evaluation and prediction condition.  [null hypothesis]  b) Comparing at the reports level, the mean SD is not different between the evaluation and prediction condition.  [null hypothesis] | a) Comparing at the adjectives level, the mean SD were not significantly different between the evaluation (*M* = 25.7) and prediction (*M* = 24.0) conditions (*t*(72) = 1.25, *p* = N/A).  b) Comparing at the reports level, the mean SD were not significantly different between the evaluation (*M* = 22.2) and prediction (*M* = 21.4) conditions (*t*(98) = 0.75, *p* = N/A). |
| 5 | *Prediction vs translation.* Three conditions:  a) Academic achievement (*n* = 32): Participants were given 10 percentile scores representing the academic achievements of different freshmen.  b) Mental concentration (*n* = 37): Participants were given 10 percentile scores representing the scores on a mental concentration test of different freshmen and told that the test is unreliable.  c) Sense of humor (*n* = 35): Participants were given 10 percentile scores representing the sense of humor of different freshmen and told that sense of humor is not an accurate predictor of GPA.  All participants were told to give their best guesses of each of the 10 students’ GPA. | a) The predictions of academic achievement and mental concentration conditions are not different. [null hypothesis]  b) The predictions between the mental concentration and sense of humor conditions are not different. [null hypothesis] | a) The mean predicted GPA between the academic achievement (*M* = 2.27) and mental concentration (*M* = 2.35) conditions were not significantly different (*p* = N/A).  b) The mean predicted GPA of the sense of humor (*M* = 2.46) was significantly higher than that in the mental concentration (*M* = 2.35) (*p* = .05). |
| 6 | *Effect of consistency:* Participants were given a student’s scores from a pair of aptitude tests that were correlated, and another student’s scores from another pair of aptitude tests that were uncorrelated. They were asked to predict the GPA for each student and report their confidence in predictions for each prediction. | Participants are more confident when predicting from the correlated tests compared to the uncorrelated tests. | Participants were significantly more confident when predicting from the correlated tests (*t*(129) = 4.80, *p* < .001) |
| 7 | *Persistence of non-regressive intuitions:* Graduate psychology students (*n* = 108) were told that a randomly selected individual scored 140 on an IQ test. Participants were asked to give the 95% CI of the true score of that individual. | Participants are more likely to give non-regressive 95% CI symmetric around 140 than regressive 95% CI. | The majority of participants gave 95% CI symmetric around 140 (73 out of 108), 24 out of 108 stated regressive 95% CI, and 11 out of 108 stated counter-regressive 95% CI. |

###### 

###### Table 2*Kahneman and Tversky (1973) Studies 1 to 7: Summary of findings*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Experiment 1 (*N* = 248)** | | | | | | |
| **Dependent Variables** |  |  |  | **Statistic** |  |  |
| Correlation between likelihood rankings and similarity rankings | | | | *r* = .97 |  |  |
| Correlation between likelihood rankings and base rates | | | | *r* = -.65 |  |  |
| **Experiment 2 (*N* = 174)** | | | | | | |
| **Dependent Variables** |  |  |  | **Statistic** |  |  |
| Law description: likelihood — similarity | | | | *r* =.93 |  |  |
| Law description: likelihood — base rates | | | | *r* = .33 |  |  |
| Computer science description: likelihood — similarity | | | | *r* = .96 |  |  |
| Computer science description: likelihood — base rates | | | | *r* = -.35 |  |  |
| Medicine description: likelihood — similarity | | | | *r* = .92 |  |  |
| Medicine description: likelihood — base rates | | | | *r* = .27 |  |  |
| Library science description: likelihood — similarity | | | | *r* = .88 |  |  |
| Library science description: likelihood — base rates | | | | *r* = .03 |  |  |
| Business administration description: likelihood — similarity | | | | *r* = .88 |  |  |
| Business administration description: likelihood — base rates | | | | *r* = .62 |  |  |
| Correlations in high accuracy vs low accuracy group | | | *t*(103) = 0.42, *p* = ns | |  |  |
| **Experiment 3 (*N* = 171)** | | | | | | |
| **Dependent Variables** |  |  |  | **Statistic** | ***df*** | ***p*** |
| Mean difference in judged probabilities in high compared to low engineer condition | | | | *t* = 3.23 | 169 | < .01 |
| **Experiment 4 (*N* = 174)** | | | | | | |
| **Dependent Variables** |  |  |  | **Statistic** | ***df*** |  |
| Adjectives: Mean difference in SD in evaluation compared to prediction condition | | | | *t* = 1.25 | 72 |  |
| Reports: Mean difference in SD in evaluation compared to prediction condition | | | | *t* = 0.75 | 98 |  |
| **Experiment 5 (*N* = 104)** | | | | | | |
| **Dependent Variables** |  |  |  | **Statistic** |  | ***p*** |
| Academic achievement — mental concentration: Mean difference in predicted GPA | | | | *t\** = N/A |  | ns |
| Academic achievement — mental concentration: Mean difference in SD of predicted GPA | | | | *t\** = N/A |  | ns |
| Academic achievement — mental concentration: Difference in slope of the regression of predicted GPA on the input scores | | | | *t\** = N/A |  | ns |
| Academic achievement — mental concentration: Difference in correlation between predicted GPA and input scores | | | | *t\** = N/A |  | ns |
| Mental concentration — sense of humor: Mean difference in predicted GPA | | | | *t\** = N/A |  | .05 |
| Mental concentration — sense of humor: Mean difference in SD of predicted GPA | | | | *t\** = N/A |  | .01 |
| Mental concentration — sense of humor: Difference in slope of the regression of predicted GPA on the input scores | | | | *t\** = N/A |  | .01 |
| Mental concentration — sense of humor: Difference in correlation between predicted GPA and input scores | | | | *t\** = N/A |  | ns |
| Academic achievement: Correlation between predicted GPA and input scores | | | | *r* = .97 |  | N/A |
| Mental concentration: Correlation between predicted GPA and input scores | | | | *r* = .95 |  | N/A |
| Sense of humor: Correlation between predicted GPA and input scores | | | | *r* = .94 |  | N/A |
| **Experiment 6 (*N* = 130\*\*)** | | | | | | |
| **Dependent Variables** |  |  |  | **Statistic** | ***df*** | ***p*** |
| Mean difference in confidence in prediction when predicting from correlated versus uncorrelated pairs of aptitude tests | | | | *t* = 4.80 | 129 | < .001 |
| **Experiment 7 (*N* = 108)** | | | | | | |
| **Dependent Variables** |  |  |  | **Statistic** |  |  |
| Proportion of participants giving non-regressive 95% CI symmetric around 140 | | | 67.59% (73/ 108) | | |  |
| Proportion of participants giving regressive 95% CI | | | 22.22% (24/ 108) | | |  |
| Proportion of participants giving counter-regressive 95% CI | | | 10.19% (11/ 108) | | |  |

*Note*. *p* values, effect sizes, and 95% CIs not listed in the table were not reported in the target article. ns = not significant. CIL = lower bounds for CIs. CIH = higher bounds of CIs. \* = [test not specified by original authors, assumed to be t-test]. \*\* inferred from *df* of t-test +1; not reported in target article

### Replication attempts

Since Kahneman and Tversky’s (1973) seminal publication on the representativeness heuristic, several direct and many conceptual replications have been conducted, with most focusing on Study 3, and relatively fewer on Study 1. Some of those studies were the basis for heated debates in the judgment and decision-making literature.

One of the closest replications of Study 3 (engineer-lawyer problem) was done by Gigerenzer et al. (1988), who gave real descriptions of engineers and lawyers to 97 participants and asked them to predict the probabilities of the randomly selected descriptions belonging to an engineer. Although they also found support for participants’ predictions deviations from Bayes’ theorem, effects were far weaker. Furthermore, Gigerenzer et al. (1988) conducted an extension in which instead of being verbally told that the description was randomly selected, participants were visually shown the process of randomly drawing a description, which further reduced the deviation from Bayesian predictions. Following this, Baratgin and Noveck (2000) successfully replicated Gigerenzer et al.'s (1988) results in a close replication. They also further examined the effect of cueing the complementarity of predictions by asking participants to also predict the probability of the given individual being from the other profession, and by telling participants that their two predicted probabilities should add up to 100% (since there are only two possible professions in the question). Baratgin and Noveck (2000) discovered that the more directly the participants were told to consider complementarity, the closer the participants’ predictions were to the Bayesian predictions.

Other less close replications of Study 3 suggested that reducing the representativeness of the descriptions could increase the use of base rates (Fischhoff & Bar-Hillel, 1984), that cueing the use of logic could increase consideration of base rates (Morsanyi & Handley, 2012), and that the Bayesian model might not accurately reflect how participants use base rates in the first place (Novemsky & Kronzon, 1999).

These replication efforts reflect the issue that the theoretical foundation and definition of the representativeness heuristic is still under controversy. In their original review paper, Kahneman and Tversky (1973) suggested that people predict by representativeness when they “select or order outcomes by the degree to which the outcomes represent the essential features of the evidence” (p. 237-238), but what constitutes essential features was not explained in detail, nor was there further theoretical elaborations on why people may tend to predict by representativeness. In a later review article, Kahneman and Frederick (2002) connected the representativeness heuristic to the dual-process model of cognitive processes (i.e., System 1 and System 2). They contended that the Tom W. study (i.e., Study 1 in the context of this paper) was an illustration of how System 2, responsible for careful and effortful thinking, was not able to correct the intuitive judgment of predicting by representativeness and not base rates when it was not prompted to do so. Kahneman and Frederick (2002) further referenced the results of the engineer-lawyer study (i.e., Study 3 in the context of this paper), in which participants predicted by representativeness despite being directly prompted about the base rates (i.e., there being 30 engineers in the described sample), and they suggested that the System 2 might be especially weak at correcting the representativeness heuristic. On the other hand, results from Gigerenzer et al. (1988) suggested that base rate neglect was much less prominent when participants were cued about the probabilistic nature of the prediction task or when participants were familiar with the context of predictions (i.e., making predictions about soccer matches). Based on these results, Gigerenzer et al. (1988) argued that base rate neglect may not be as general to human thinking as originally conceptualized, but rather depended on the presentation of the judgment problem. In response, Kahneman and Frederick (2002) maintained that the studies showing the conditions in which people do not predict by representativeness were simply illustrating the limitations of System 2 but not the conceptualization of the representativeness heuristic. It is evident that the definition and underlying theory for the representativeness heuristic has been subject to heated debate.

Finally, Koehler (1996) raised some doubts regarding representativeness heuristic. Reviewing work building on Kahneman and Tversky (1973), Koehler (1996) argued that the effect of the representativeness heuristic is not large enough to clearly conclude that base rates are largely ignored, and that variations in the presentation of the task (e.g., showing that the sampling of descriptions is random and repeated) may further reduce the effect.

In summary, Kahneman and Tversky (1973) have been extremely influential, yet have also been the subject of some debate with mixed findings and doubts regarding robustness, effect size, and interpretation. Much of the literature has focused on the paradigm in Study 3, yet - as far as we know[[1]](#footnote-2) - with no comprehensive replication of all seven studies described in Kahneman and Tversky (1973) together looking at the effects systematically.

## Choice of article for replication: Kahneman and Tversky (1973)

We embarked on a Replication Registered Report of Kahneman and Tversky (1973) with several adjustments and added extensions. We opted to include all seven studies in the target article. We aimed to revisit the phenomenon to examine the reproducibility and replicability of the findings with an independent pre-registered well-powered replication and extension. This follows the recent growing recognition of the importance of reproducibility and replicability in psychological science (e.g., Nosek et al., 2022; Zwaan et al., 2018).

We chose Kahneman and Tversky (1973) based on several factors: the potential for improvements in methodology and/or theory, its academic and practical impact, potential for further extensions to examine the effect of consistency in numerical predictions and that of statistical knowledge on the persistence of the heuristic with better methods, and the absence of direct replications.

The article had an extraordinary impact on scholarly research in the areas of social psychology, judgment and decision making, behavioral economics, law, policy, and many other fields. At the time of writing (January, 2024), there were 9803 Google Scholar citations of the article and a few important follow-up theoretical and empirical articles. For example, Gigerenzer et al. (1988) engaged in a heated debate regarding the results of Kahneman and Tversky (1973) about the very nature and interpretation of heuristics and biases, and Koehler (1996) commented on the theoretical and practical problems of the effect size and applicability of the representativeness heuristic.

We consider conducting a comprehensive replication of many similar studies reviewed in a single review article as an important and promising research direction. We previously conducted systematic replications of seminal review papers (Au & Feldman, 2020; Hong & Feldman, 2024; Li & Feldman, 2024) and of papers reviewing a large number of empirical studies on a single phenomenon (Mayiwar et al., 2024 replication of a different article by Kahneman and Tversky, 1972). We see much value in those, especially as for the target phenomenon they allow for: 1) a comprehensive empirical test of many different methods, and 2) the examination of in-person consistency in responding to many problems and scenarios.

The design of the studies in this paper allowed for the straightforward inclusion of extensions to allow for additional tests and insights on Study 4 and Study 7 of the target article, which respectively examined how confidence in predictions may be related to numerical predictions, and how an individual’s statistical knowledge affect the persistence of non-regressive intuitions. We observed that the original measures of certainty in predictions in Study 4 can be tested in a more direct way, and that the study design of Study 7 can be improved to reduce the influence of potential artifacts. More detailed descriptions of the extensions can be found further down in the extension section.

When we analyzed the article we realized that some analyses were in need of revisiting and improvement. The analyses in Study 3, 4, 5, and 7 in the original were reliant on null effects that were attempted to be substantiated through nonsignificant *p* values in t-tests. In this replication, we aimed to improve on those by first replicating the testing of null hypotheses using Null Hypothesis Significance Testing (NHST), reframing null hypotheses to their alternatives, and if we indeed find that we fail to reject the null and find support for the alternative, then we will supplement those with Bayesian statistics to quantify support for the null. Furthermore, Study 2 of the target article involved deception regarding participants’ perceived predictive accuracy, which we deemed unnecessary and unconvincing. Our replication modified the study design such that participants self-reported their perceived predictive accuracy instead such that a correlational analysis can be run without the need for deception.

Despite its impact, to our knowledge, there are currently only a few published direct replications of only a few of the studies reviewed. We therefore aimed to revisit the classic phenomenon to examine the reproducibility and replicability of the phenomenon with an independent well-powered close replication and extension Registered Report of Kahneman and Tversky (1973) .

## Extensions

We aimed to extend the replication study by considering more straightforward measures of certainty in Study 4 and improving the study design of Study 7 with correlation analyses to reduce artifacts and investigate the strength of the effect in greater detail.

### Study 4: Directly measuring confidence in predictions

In Study 4 of the target article, Kahneman and Tversky (1973) theorized that when given information of different individuals to predict from, people would be more certain about the accuracy of their answers if people made predictions based on representativeness, which would be reflected in a larger standard deviation within their answers. In other words, participants would see that descriptions of each individual are representative of a distinctly different characteristic, which would give them higher certainty in making varied predictions for each of them. In our extension, in addition to replicating the SD analyses, we directly asked the participants to report their confidence in their predictions, which we theorized would give a more straightforward estimate of participants’ confidence due to predicting by representativeness.

### Study 7: Correlational analysis of regressive intuitions and statistical knowledge

For Study 7, the target article recruited psychology graduate students who supposedly should have prior knowledge of statistics and regression, then asked them a statistical problem which was meant to test their ability to consider regression to the mean, which most failed in. We observed that this design might have left room for artifacts such as whether the students were actually adept at psychology statistics, whether the university taught statistics well, etc. As such, we modified the study design where participants would be asked the same original problem but then would self-report their level of statistical knowledge instead. This design can alleviate the ecological issues of the original and would allow for a correlation analysis which can better capture the size and direction of the effect.

From a theoretical standpoint, the relationship between statistical knowledge and non-regressive intuitions might seem slightly removed from the phenomenon of the representativeness heuristic. However, Kahneman and Frederick (2002) reported that statistical expertise did not help avoid predicting by representativeness unless prompted. However, there is some evidence that numeracy moderated some of the classic effects in decision-making. In four studies, Peters et al. (2006) showed that numeracy moderated four classic effects (e.g., framing effects), and our replication found similar yet weaker support for three of those (Zhu & Feldman, 2023), showing that numeracy does not always lead to better evaluations, and sometimes tends to back-fire. We added an extension to Study 7 to examine the association between statistical knowledge and the effect.

## Overview of replication and extension

Kahneman and Tversky’s (1973) empirical work consisted of seven experiments, and we aimed to include all seven. We combined Studies 1 and 2 due to their similar design and made a slight adjustment to the design to avoid using deception. Participants were randomly assigned to one of three groups to give the base rates of nine fields of specialization, to rank the nine fields in order of similarity with five described individuals, or to predict the likelihood rankings of these five individuals being a student in the nine fields. Participants in the likelihood group also reported their self-perceived accuracy in predictions. We tested the correlations between the likelihood rankings, similarity rankings, base rates, and self-perceived accuracy.

In Study 3, participants were given descriptions of individuals supposedly randomly drawn from a sample of engineers and lawyers. Half of the participants were told the sample consisted of 30 engineers and 70 lawyers, and another half was told the sample was 70 engineers and 30 lawyers. We tested the differences in the mean probability judgements between the groups and whether predictions fit the curved line of the correct relation according to Bayes’ rule.

In Study 4, participants were given either adjectives or reports (between-subjects) about freshmen and asked to make either evaluations or predictions of their academic ability. We tested the differences in SD between the four conditions and extended the original study to test the differences in confidence in estimations between the conditions.

In Study 5, participants were given input scores supposedly representing academic achievement, mental concentration, or sense of humor of ten students (between-subjects), then they gave predictions about their GPA. We tested the differences in mean predicted GPA, SD, and correlation between the conditions.

In Study 6, participants were given two pairs of aptitude tests in which one pair was correlated and the other was not. Participants predicted the corresponding student’s GPA for each pair of tests and reported their confidence in prediction. We tested the mean difference in confidence in predictions between the two pairs of tests.

In Study 7, participants were asked about a statistical problem regarding confidence intervals. Participants answered with a 95% CI and self-reported their levels of statistical knowledge. We replicated the descriptives of the proportions of participants giving non-regressive, regressive, and counter-regressive confidence intervals. We extended the original study to test the correlation between participants’ level of statistical knowledge and the degree of regression in their answers.

## Deviations

We made a few adjustments with reference to the target article design, summarized in Table 3.

###### Table 3 *Replication and extension adjustments to the target article’s methods and design*

| **Studies** | **Factor** | **Target article** | **Adjustment in current study** | **Reason for change / Justifications** |
| --- | --- | --- | --- | --- |
| 1 and 2 | Study design | Studies 1 and 2 were originally carried out separately with different samples. | Studies 1 and 2 were combined to be run with one single sample and random allocation between conditions. | The experimental design of Studies 1 and 2 were extremely similar (comparing base rates, similarity, and likelihood rankings). |
| 2 | Study design | All participants in the similarity and prediction conditions read all personality sketches. | In our combined Studies 1 and 2, participants in the similarity and likelihood conditions only read one out of six personality sketches. | Concerns of participant fatigue due to the combined study design of running all seven studies with one sample. |
| 2 | Conditions, DV | Participants in the prediction condition were further split between the high and low accuracy conditions. | Participants were not split between high and low accuracy conditions, instead all participants in the prediction condition were asked to report their self-perceived accuracy. | Target article involved deceiving participants about their predictive accuracy, which we considered unnecessary and unconvincing. Furthermore, the original authors reported that the manipulation did affect the participants' prediction. In this case, modifying the experimental manipulation into a correlational approach may help detect the difference in greater detail. |
| 2 | Materials | Participants were told that the descriptions shown were thumbnail personality sketches of ninth-grade boys allegedly written by counselors (authors did not clarify whether these were artificially created). | Personality sketches used for the hypothetical individuals that were implied to be in business administration, computer science, law, library science, and medicine were generated by ChatGPT (detailed in the supplementary). The engineering (Tom W.) description was from the target article. Participants were not told about how the descriptions were obtained. | Lack of documentation in the target article and difficulties in finding the original materials in subsequent replications. |
| 3 | Study design | All participants in the high and low engineer conditions were shown all six descriptions. | Participants in the high and low engineer conditions only read one out of six descriptions. | Concerns of participant fatigue due to the combined study design of running all seven studies with one sample. |
| 3 | Procedural details | Target article included the prompt “The same task has been performed by a panel of experts, who were highly accurate in assigning probabilities to the various descriptions. You will be paid a bonus to the extent that your estimates come close to those of the expert panel.” | Removed in our replication. | The original implementation was very unclear, leaving too many questions open. We did not consider this to be core to the hypotheses or the empirical demonstration and therefore instead opted for a simple execution with no reference to that panel or to issuing bonuses. |
| 3 | Materials | Original authors reused the Tom W. description from Study 1 as an engineer description for Study 3. | We replaced the Tom W. description with another engineer description generated by ChatGPT (detailed in the supplementary). Moreover, we added one lawyer description referenced by Gigerenzer et al. (1988). | For the reused Tom W. description, the concern was of artifacts due to familiarity with the questions from Studies 1 and 2 since all seven studies were run with one sample. For the lawyer's description, replacement was due to lack of documentation in the target article and difficulties in finding the original materials in subsequent replications. |
| 4 | Study design | All participants in the adjectives and reports conditions were shown all descriptions. | Participants in the adjectives and reports conditions were only shown one out of nine descriptions. | Concerns of participant fatigue due to the combined study design of running all seven studies with one sample. |
| 4 | Procedures | Participants in the evaluation condition were asked to estimate “the percentage of students in the entire class whose descriptions indicate a higher academic ability.” | The prompt was modified to “estimate, in terms of a percentile score, how impressive the described student is to you in terms of academic ability compared to the other students in the entire class (Example: a percentile score of 65 means that you consider this student to be more impressive in terms of academic ability than 65% of his class)” | The original prompt seemed to be counter-intuitive in direction compared to the plots of the original results, with a lower percentile score implying being more impressed. |
| 4 | Materials | Original descriptions of adjectives and reports were used. | We generated eight additional adjective descriptions consisting of five adjectives and nine reports descriptions using ChatGPT based on the only original adjective description available in the target article: “intelligent, self-confident, well-read, hard- working, and inquisitive” (p. 243) | Lack of documentation in the target article and difficulties in finding the original materials in subsequent replications. |
| 4 | Dependent variable | A direct measure of confidence in estimations was not included. | We included a direct self-reported measure of confidence in estimations as an extension. | This could more directly measure the concept of certainty in predictions compared to the indirect approach of inferring through SD in predictions in the target article. |
| 5 | Procedures | The original prompt for the GPA prediction was “... give your best guess about his grade point average for *that* year” | The prompt was modified to “... give your best guess about his grade point average for the *second* year” (the changed word is italicized). | The original phrasing being ambiguous and potentially making participants confused about using scores on academic achievement to predict GPA in the same year. |
| 6 | Materials | The scores used for the correlated and uncorrelated pairs of aptitude tests were not specified. | For the correlated pairs of tests, the scores were 80 and 89; for the uncorrelated pairs of tests, the scores were 80 and 63. | Lack of documentation in the target article and difficulties in finding the original materials in subsequent replications. |
| 7 | Sample | The sample was graduate students in psychology affiliated with the authors. | A sample of Americans recruited through Prolific. | Avoiding artifacts, and practical limitations. |
| 7 | Procedures | The prompt of the target article did not explain the concept of IQ scores except that “... an IQ score is the sum of a "true" score and a random error of measurement which is normally distributed” (Kahneman & Tversky, 1973, p. 250). | The sentence “The average IQ test score is 100, and 68% of all people are between 85 and 115” was added in the prompt. | Some participants might not be familiar with the statistical properties of IQ scores. |
| 7 | DV | A measure of statistical knowledge was not included. | Three extension measurements of participants’ self-reported level of statistical knowledge, usage, and training were included. | This could better capture the effect of prior statistical knowledge on the persistence of non-regressive intuitions whilst avoiding the potential artifacts in the original design. |

## Pre-registration and open-science

We provided all materials, data, and code on: <https://osf.io/8zhcj/>. This project received Peer Community in Registered Report Stage 1 in-principle acceptance ((Enter link after IPA); (Enter link after IPA)) after which we created a frozen pre-registration version of the entire Stage 1 packet (Enter link after IPA) and proceeded to data collection. All measures, manipulations, exclusions conducted for this investigation are reported, and data collection was completed before analyses. This Registered Report was written using the Registered Report template by Feldman (2023).

# Method: All studies

[IMPORTANT: Method and results were written using a randomized dataset produced by Qualtrics to simulate what these sections will look like after data collection. These will be updated following the data collection. For the purpose of the simulation, we wrote things in past tense, but no pre-registration or data collection took place yet.]

This section reported the methodology that applied to all six studies. The exact details regarding design, manipulations, measures, and corresponding results for each of the studies were reported in the subsequent sections separately.

## Power and sensitivity analyses

Effect sizes, confidence intervals, and power were calculated with the help of a guide by Jané et al. (2024) and R (Version 4.3.2; R Core team, 2023) using packages “pwr” (version 1.3-0) for the factors that the authors found support for in the target article (flagged as significant results).

We attempted to calculate effect sizes (ES) based on the effects reported in the target article, yet in most cases the target article did not report effect sizes, utilized tests that were not compatible with participant-level power analyses (e.g., item-level correlations in Studies 1 and 2), and aimed to test null effects with NHST (perfectly understandable given the times) rather than aiming to reject the null hypothesis or using more appropriate methods (we discuss these challenges further in the “Effect size calculations of the target article effects” section of the supplementary). We concluded that this made conducting power analyses either relatively unhelpful or difficult.

On Prolific, we set to collect a sample of 800[[2]](#footnote-3) participants (200 participants per experimental condition in Study 4, which has the greatest number of conditions), which is over three times larger than the samples reported in the target article. We conducted a sensitivity analysis using Gpower (Faul et al., 2007), which indicated that a sample of 800 with a target alpha of .005, we could detect *f* = 0.16 for the 2 by 6 ANOVA in Study 3, *f* = 0.16 for the 2 by 2 ANOVA in Study 4, *d* = 0.39 for the independent samples t-test in Study 5, *d* = 0.16 for the paired samples t-test in Study 6, and *r* = 0.16 for the bivariate correlation in Study 7. These are commonly considered small to moderate effects in social psychology (Jané et al., 2024). As a reminder, to allow for an easy comparison, the target article’s Study 1 had 248 participants, Study 2 had 174 participants, Study 3 had 171 participants, Study 4 had 174 participants, Study 5 had 104 participants, Study 6 had 130 participants, and Study 7 had 108 participants. We provided more information regarding the difficulties with calculations of effect sizes and power in the “Power analysis” subsection of the supplementary materials.

## Participants

[To demonstrate what the results would look like after data collection we simulated a dataset of 1309 participants using Qualtrics and reported our analyses based on that dataset. Results will later be updated in full to a sample of 800 and the real data.]

We recruited a total of 1309 US Americans on Prolific (Palan & Schitter, 2018; *Mage* = 49.83, *SD* = 28.33; 330 females, 325 males, 654 other or did not disclose). Participants were 18 years old and above and were born, raised, and residing in the US. We targeted US Americans using Prolific’s filters. We restricted the location to the US using “standard sample”, we set it to “Nationality: United States”, “Country of birth: United States”, “Minimum Approval Rate: 90, Maximum Approval Rate: 100”, “Minimum Submissions: 50, Maximum Submissions: 100000”. We note that [TBD] subjects began the survey but [TBD] did not proceed beyond the consent and verifications. We summarized a comparison of the target article sample and the replication samples in Table 4.

[Stage 1 note: We will first pretest the survey duration and technical feedback with 30 participants to make sure our time run estimate was accurate and adjusted pay as needed, the data of the 30 participants will not be analyzed other than to assess survey completion duration, feedback regarding possible technical issues and payment, and needed pay adjustments. Unless in the case of serious technical issues that affect data quality and require survey modification, these participants will be included in the overall analyses.]

[The assignment pay is based on the federal wage of 7.25USD/hour, per minute, so for example 5-8 minutes survey would be paid 1 USD per participant. We first pretested survey duration with 30 participants to make sure our time run estimate was accurate and adjusted pay as needed, the data of the 30 participants was not analyzed other than to assess survey completion duration and needed pay adjustments. For those pretest participants, if survey duration was longer than expected, they were paid a bonus as pay adjustment. The pretest participants' responses were included in the final analysis.]

###### 

###### Table 4 *Difference and similarities between original study and replication*

|  |  |  |
| --- | --- | --- |
|  | Kahneman and Tversky (1973) | US on Prolific |
| Sample size | Study 1: 248 Study 2: 174 Study 3: 171 Study 4: 174 Study 5: 104 Study 6: 130 Study 7: 108 | 1309 | |
| Geographic origin | Unclear, but those specified were US American[[3]](#footnote-4), and given authors’ base, possibly Israeli samples. | US American | |
| Gender | N/A | 325 males, 330 females, 654 other/did not disclose | |
| Median age (years) | N/A | 50 | |
| Average age (years) | N/A | 49.83 | |
| Standard deviation age (years) | N/A | 28.33 | |
| Age range (years) | N/A | 0-100 | |
| Medium (location) | N/A | Computer (online) | |
| Compensation | Payment (amount not specified) | Nominal payment | |
| Year | 1973 or before | 2024 | |

*Note*. N/A = unknown, not provided.

### 

### Design: Replication and Extension

In the target article, the studies were all conducted separately with independent samples, some with more than a single sample in the different conditions. We ran the seven studies together in a single unified data collection. The display of scenarios and conditions was counterbalanced using the randomizer “evenly present” function in Qualtrics. This unified design combining replications of several studies into a singular data collection was previously tested successfully in many of the replications and extensions conducted by our team (e.g., Petrov et al., 2023; Vonasch et al., 2023; Yeung & Feldman, 2022; Zhu & Feldman, 2023), and is especially powerful in addressing concerns about the target sample (e.g., naivety and attentiveness) when some studies replicate successfully whereas others do not, as well as in allowing for drawing inferences about links between the different studies and consistency in participants’ responding to similar decision-making paradigms.

[Note: We will test for order effects, with each study when it is displayed first. See “data analysis strategy” section.]

## Procedure

[*For review: The Qualtrics survey .QSF file and an exported DOCX file are provided on the OSF folder. A preview link of the Qualtrics survey is provided on:*  
[<https://hku.au1.qualtrics.com/jfe/preview/previewId/fcd268ef-ea89-46d4-9cf6-50cbb78a3df5/SV_eznJKe4H21mPy8S?Q_CHL=preview&Q_SurveyVersionID=current> ]

We reconstructed the target’s stimuli and adjusted it to an online Qualtrics survey based on the information provided in the article. Participants indicated their consent, with questions confirming their eligibility, understanding, agreement with study terms, and ability to comprehend and attend to the survey, which they must answer with a “yes” and required responses in order to proceed to the study. Three of the four questions also served as attention checks, with the order of the options being rotated (yes, no, not sure). Participants completed the six studies of the combined Studies 1 and 2, and Studies 3 to 7, in random presentation order. Afterwards, participants reported their level of statistics knowledge.

At the end of the experiment, participants answered a number of funneling and demographic questions, and were debriefed.

## Evaluation criteria for replication findings

Although the effect sizes for the original Study 3 to 7 were not reported, the item-level correlations for Studies 1 and 2 were reported. As such, we aimed to compare the replication effects with the original effects in the target article using the criteria set by LeBel et al. (2019) (see section “Replication evaluation” in the supplementary).

We pre-registered our overall strategy to conclude a successful replication if at least 4 out of the 6 studies showed a signal in the same direction as the target article, a failed replication if only 0 or 1 studies, and mixed findings if 2 or 3 studies.

## Replication closeness evaluation

We provided details on the classification of the replications using the criteria by LeBel et al. (2018) criteria in Table 5. We summarized the replication as a close replication for Studies 3 to 7, and a conceptual replication for Studies 1 and 2.

###### Table 5 *Classification of the replication based on LeBel et al. (2018)*

|  |  |  |
| --- | --- | --- |
| **Design facet** | **Replication** | **Details of deviation and severity** |
| Effect/hypothesis | Same |  |
| IV construct | Same |  |
| DV construct | Same |  |
| IV operationalization | Similar | Study 2: We did not implement the deception of accuracy, and instead changed to a continuous measure testing for a correlation.  Studies 1-2: Combined design, changes to the studies and analyses to fit combined design.  Study 7: Statistical knowledge was measured as a continuous self-reported individual differences factor. |
| DV operationalization | Same |  |
| IV stimuli | Similar / Unknown | Due to lack of original documentation, we attempted to replicate the materials used for Study 1, 2, 3, 4, and 6 by creating new stimuli with the help of ChatGPT. |
| DV stimuli | Similar | The wordings for the DV of Study 4 and 5 were adjusted to provide better clarity. |
| Procedural details | Different | All seven studies were run with one sample. Task instructions were adjusted to improve clarity. |
| Physical settings | Different | All participants participated online using a computer. |
| Contextual variables | Different | It has been 50 years since the original study was published. |
| Population | Unknown / Similar | Population characteristics in the original were lacking except for the note that “subjects in the studies reported in this paper were paid volunteers recruited through a student paper at the University of Oregon.” (Kahneman & Tversky, 1973, p. 238). |
| Replication classification | Studies 1-2: Far replication.  Studies 3-7: Close replication | |

*Note*. Criteria for evaluation of replications by LeBel et al. (2018). "Similar" category was added to the LeBel et al. (2018) typology to refer to minor deviations or extensions aimed to adjust the study to the target sample that are not expected to have major implications on replication success.

## Data analysis strategy

We set our target alpha to .005 for all analyses. The decision was aiming to account for the unified design of seven studies with some studies having multiple analyses. We performed the analyses for all studies with R (Version: 4.3.2) with packages “ggplot2” (Version: 3.5.0), “ggstatsplot” (Version: 0.12.2), “haven” (Version: 2.5.4), “knitr” (Version: 1.45), “effectsize” (Version: 0.8.6), “jmv” (Version: 2.4.11), “dplyr” (Version: 1.1.4), “car” (Version: 3.1.2), “remotes” (Version: 2.4.2.1), “varBF” (Version: 0.0.1), and “grateful” (Version: 0.2.4) (see provided full Rmarkdown code based on simulated data in the OSF folder). The Bayesian analyses for Studies 3, 4 (extension), and 5 were performed with the “jsq” module on JAMOVI.

### Outliers and exclusions

We did not classify outliers in this study. All data from participants who successfully completed the survey were included.

### Order effects

One deviation from the target article is that we combined all studies to a unified design such that all participants complete all studies/scenarios in one setting in random order. We considered this to be a stronger design with many advantages, yet one disadvantage is that answers to one scenario may bias participants’ answers to following scenarios.

We, therefore, pre-registered that we would examine order as an exploratory moderator, meaning that we will run the analyses first with the study displayed first and then with the study not displayed first, and report the differences between the two, and examine whether the confidence intervals of the effect overlap. To compensate for multiple comparisons and the increased likelihood of capitalizing on chance, we set the alpha for the additional analyses to a stricter .001.

[TBD conclusion based on our experience with a unified design so far: We found [no] differences in conclusions]

### Bayesian analyses

We pre-registered that in case we failed to find support for the hypothesis for any of the studies, we would run a complementary Bayesian analysis for that study using a prior of 0.707 to quantify support for the null.

# Studies 1 and 2: Similarity and predictions

## Design

We combined Studies 1 and 2 into a unified between-subjects design contrasting three conditions: base rates, similarity, and likelihood. We summarized the design in Table 6.

###### 

###### Table 6 *Studies 1 and 2 (similarity and predictions): Experimental design [Between subject]*

|  |  |  |
| --- | --- | --- |
| **Base rates condition**  Participants estimated the percentage of current first year graduate students enrolled in nine fields of specialization:  Business administration  Computer science; Engineering; Humanities and education; Law; Library science; Medicine; Physical and life sciences; Social sciences and social work. | **Similarity condition**  Participants were given a personality sketch of an individual which is implied to be either an engineering, a computer science, law, medicine, library science, or business administration student (random assignment; between subjects) and asked to rank how similar he is to the typical graduate student in the nine fields of specialization. | **Predictions condition**  Participants were presented with personality sketches of individuals fitting stereotypes of engineering, computer science, law, medicine, library science, or business administration students (random assignment; between subjects) and also a null description.  They ranked the nine fields of specialization in order of likelihood the person is a graduate from that field. |
| Dependent variable: [Studies 1 and 2]  **Judged base rates**  Scale:  0% = *None enrolled*;  100% = *All enrolled*  (all nine must add up to 100%) | Dependent variable:  [Study 1]  **Similarity ranking**  Participants rank similarity.  Scale:  1 = *Most similar*; 9 = *Least similar*  (Qualtrics ranking question type) | Dependent variables:  [Study 2]  **Likelihood ranking**  Participants rank likelihood. Scale:  1 = *Most likely*; 9 = *Least likely*  **Accuracy estimation:**  Participants estimated the probability accuracy.  Scale: 0% to 100% |

## Manipulations

In Studies 1 and 2, participants were evenly and randomly assigned into one of three conditions: base rate, similarity, or prediction.

In the base rate condition, participants were asked to guess the percentage of current first year graduate students enrolled in nine fields of specialization (e.g., business administration, computer science, engineering, etc.). In the similarity condition, participants were further evenly and randomly split into groups to be shown either the business administration, computer science, engineering, law, library science, or medicine description. They were then asked to rank how similar the described individual is to the typical graduate student in the nine fields of specialization. In the prediction condition, participants were similarly split into groups to be shown one of the six aforementioned descriptions, but were instead asked to rank the nine fields of specialization on the likelihood of the described individual being a graduate from that field. The prediction condition participants were further given a null description which provided no information about an individual Don T. except that he is a first-year graduate student and were asked to make the same likelihood rankings.

## Measures

### Judged base rates

Participants reported their estimates of the percentage of current first year graduate students enrolled in the nine fields of specialization of business administration, computer science, engineering, humanities and education, law, library science, medicine, physical and life sciences, and social science and social work (0% = *None enrolled*, 100% = *All enrolled*).

### Similarity ranking

Participants ranked the nine fields of specialization in terms of similarity between the individual description they were given and the typical student in that field. (1 = *Most similar*, 9 = *Least similar*).

### Likelihood ranking

Participants ranked the nine fields of specialization in terms of the likelihood that the individual described is enrolled in that field. (1 = *Most likely*, 9 = *Least likely*).

### Self-perceived accuracy

We modified the study design to incorporate the manipulation of perceived accuracy in the original Study 2 as a continuous measure of participants’ self-perceived accuracy. Specifically, participants reported the probability that they made correct predictions about the inquired likelihood for the individual shown (0% to 100% certainty).

## Data analysis strategy

### Replication: As in the original

Mirroring the target article, we ran the following main analyses: a) For each of the six descriptions (business administration, computer science, engineering, library science, law, and medicine), we calculate the item-level correlation between the mean likelihood ranking and mean base rates for each of the nine fields of specialization; b) for each of the six descriptions, we calculate the item-level correlation between the mean likelihood ranking and mean similarity ranking for each of the nine fields of specialization. Deviating from the target article due to our combined study design, to address the same research question of the effect of participants’ self-perceived predictive accuracy, we calculated c) item-level correlations between the mean self-perceived accuracy in the six descriptions and the *r* value of likelihood-base rates for the six descriptions. Finally, to match the target article, we calculated d) the correlation between mean base rates and mean likelihood rankings for the null description (the null was shown to all participants in the prediction condition).

The target article did not report correlations between the base rates and the similarity rankings, likely because the question of interest was regarding whether individuals’ predictions (i.e., likelihood rankings) were more associated with prior probabilities (i.e., base rates) or with the representativeness of available evidence (i.e., similarity rankings). Yet, we thought it worthwhile to also test for the correlations between base rates and similarity rankings to see whether participants’ perception of representativeness matched up with the base rates as well. Therefore, we also conducted an additional analysis and for each of the six descriptions, we calculated e) the item-level correlation between the mean similarity ranking and mean base rates for each of the nine fields of specialization.

## Results and discussion

[Reminder for Stage 2: Alpha is set to .005]

We summarized the mean estimated base rates, similarity rankings, and likelihood rankings for the nine areas of specialization in Table 7, the findings for Studies 1 and 2 in Table 8, and the mean reported self-perceived accuracy in predictions in the six descriptions in Table 9.

[Note for reviewers: We observed that these mean values were skewed towards the fields of specialization that were presented earlier (e.g., mean base rates in the fields of business administration and computer science), which was mostly likely due to quirks in Qualtrics’ generated test responses for responses that most sum to 100 (50-50 for the first, 25-25 for the second, etc.). ]

###### 

###### Table 7 *Studies 1 and 2: Means of base rates, similarity rankings, and likelihood rankings in nine areas of specialization*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Base** | **Similarity** | | | | | | **Prediction** | | | | | | |
|  | **Base** | **Business** | **Computer science** | **Engineering** | **Law** | **Library**  **science** | **Medicine** | **Business** | **Computer science** | **Engineering** | **Law** | **Library**  **science** | **Medicine** | **Null** |
| ***n*** | 436 | 73 | 72 | 73 | 73 | 73 | 73 | 73 | 73 | 72 | 73 | 72 | 73 | 436 |
| Business administration | 50.08 | 5.25 | 4.50 | 5.40 | 5.30 | 4.82 | 5.21 | 4.80 | 5.15 | 5.54 | 5.33 | 4.74 | 4.85 | 4.84 |
| Computer science | 25.30 | 4.69 | 5.46 | 5.25 | 4.58 | 4.55 | 5.23 | 5.26 | 4.93 | 4.71 | 4.66 | 5.00 | 5.19 | 4.95 |
| Engineering | 11.80 | 5.10 | 4.82 | 5.44 | 5.04 | 5.03 | 4.92 | 4.74 | 5.07 | 4.99 | 5.21 | 5.04 | 4.99 | 5.01 |
| Humanities and education | 6.55 | 4.81 | 5.24 | 4.62 | 5.48 | 5.63 | 4.86 | 5.33 | 4.99 | 4.94 | 5.34 | 5.14 | 5.12 | 5.04 |
| Law | 3.17 | 4.92 | 4.93 | 4.23 | 4.85 | 4.80 | 4.77 | 4.85 | 4.25 | 5.07 | 5.01 | 4.64 | 4.75 | 5.16 |
| Library science | 1.62 | 5.01 | 5.40 | 5.25 | 5.10 | 5.16 | 4.52 | 5.21 | 4.84 | 4.74 | 5.10 | 5.21 | 4.93 | 5.11 |
| Medicine | 0.75 | 5.03 | 5.08 | 5.03 | 4.82 | 5.01 | 4.90 | 5.58 | 5.22 | 4.88 | 5.06 | 5.31 | 4.82 | 4.70 |
| Physical and life sciences | 0.39 | 5.07 | 4.74 | 4.78 | 4.73 | 4.97 | 5.73 | 5.11 | 5.34 | 5.46 | 4.30 | 5.00 | 5.44 | 4.94 |
| Social science and social work | 0.34 | 5.14 | 4.83 | 5.01 | 5.11 | 5.03 | 4.86 | 4.14 | 5.22 | 4.68 | 5.00 | 4.93 | 4.90 | 5.26 |

*Note*. The null description was shown to all participants in the likelihood prediction condition.

###### Table 8 *Studies 1 and 2: Summary of statistical tests, effects, and evaluation*

|  |  | **Replication** | | | **Target article** |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Analysis** | **DV** | ***r*** | ***p*** | **95% *CI*** | **Effect** | **Interpretation** |
| **a)** | Business administration description: likelihood — base rates | -.081 | .8 | [-.707, .616] | *r* = .62 | no signal — inconsistent |
|  | Computer science description: likelihood — base rates | .098 | .8 | [-.605, .716] | *r* = -.35 | no signal — inconsistent |
|  | Engineering description: likelihood — base rates | .454 | .2 | [-.301, .859] | *r* = -.65 | no signal — inconsistent |
|  | Law description: likelihood — base rates | .280 | .5 | [-.472, .796] | *r* = .33 | no signal — consistent |
|  | Library science description: likelihood — base rates | -.436 | .2 | [-.853, .321] | *r* = -.03 | no signal — consistent |
|  | Medicine description: likelihood — base rates | .102 | .8 | [-.717, .603] | *r* = .27 | no signal — inconsistent |
| **b)** | Business administration description: likelihood — similarity | -.541 | .2 | [-.878, .228] | *r* = .88 | no signal — inconsistent |
|  | Computer science description: likelihood — similarity | -.295 | .4 | [-.802, .459] | *r* = .96 | no signal — consistent |
|  | Engineering description: likelihood — similarity | -.069 | .9 | [-.701, .624] | *r* = .97 | no signal — consistent |
|  | Law description: likelihood — similarity | .797 | .010 | [.282, .955] | *r* = .93 | signal — inconsistent, smaller |
|  | Library science description: likelihood — similarity | .487 | .2 | [-.261, .870] | *r* = .88 | no signal — inconsistent |
|  | Medicine description: likelihood — similarity | .729 | .026 | [.126, .939] | *r* = .92 | signal — inconsistent, smaller |
| **c)** | Self-perceived accuracy — likelihood-base rate correlation coefficients | .464\* | .4 | [-.557, .927] | null effect | no signal — consistent |
| **d)** | Null description: likelihood — base rates | -.370 | .3 | [-.830, .390] | *r* = .74 | no signal — inconsistent |
| **f)** | Business administration description: similarity — base rates | .179 | .6 | [-.551, .753] |  |  |
|  | Computer science description: similarity — base rates | -.320 | .401 | [-.812, .437] |  |  |
|  | Engineering description: similarity — base rates | .504 | .166 | [-.240, .875] |  |  |
|  | Law description: similarity — base rates | .195 | .615 | [-.539, .761] |  |  |
|  | Library science description: similarity — base rates | -.410 | .272 | [-.844, .349] |  |  |
|  | Medicine description: similarity — base rates | .280 | .466 | [-.472, .796] |  |  |

*Note*. Pearson’s correlation, *df* = 7. *CI* = 95% confidence intervals. The interpretation of outcome was based on LeBel et al. (2019). \*: *df* = 4

###### Table 9 *Studies 1 and 2: Mean reported self-perceived accuracy in predictions*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Business administration** | **Computer**  **science** | **Engineering** | **Law** | **Library**  **science** | **Medicine** |
| 50.85 | 47.77 | 52.43 | 50.99 | 49.78 | 49.9 |

#### a) Item-level correlations between the mean likelihood ranking and mean base rates for each of the nine fields of specialization using the six descriptions

We calculated Pearson's correlation between the mean base rates and the mean likelihood ranking across the nine fields of specialization at the item level, repeated for each of the six descriptions shown to the participants (i.e., business administration, computer science, engineering, law, library science, medicine). For the analysis, the likelihood ranking was reversed such that a positive correlation would mean a high likelihood ranking being associated with high estimated base rates.

We found no support for a correlation between likelihood rankings and base rates in the business administration description (*r*(7) = -.081, *p* = .8, 95% CI [-.707, .616]), computer science description (*r*(7) = .098, *p* = .8, 95% CI [-.605, .716]), engineering description (*r*(7) = .454, *p* = .2, 95% CI [-.301, .859]), law description (*r*(7) = .280, *p* = .5, 95% CI [-.472, .796]), library science description (*r*(7) = -.436, *p* = .2, 95% CI [-.853, .321]), and medicine description (*r*(7) = -.102, *p* = .8, 95% CI [-.717, .603]).

#### b) Item-level correlations between the mean likelihood ranking and mean similarity ranking for each of the nine fields of specialization using the six descriptions

We calculated Pearson's correlation between the mean likelihood ranking and the mean similarity ranking across the nine fields of specialization at the item level, repeated for each of the six descriptions shown to the participants. We found support for a positive correlation between likelihood and similarity ranking for the law description (*r*(7) = .797, *p* = .010, 95% CI [.282, .955]), and the medicine description (*r*(7) = .729, *p* = .026, 95% CI [.126, .939]). We found no support for a correlation between likelihood rankings and base rates in the business administration description (*r*(7) = -.541, *p* = .2, 95% CI [-.878, .228]), computer science description (*r*(7) = -.295, *p* = .4, 95% CI [-.802, .459]), engineering description (*r*(7) = -.069, *p* = .9, 95% CI [-.701, .624]), library science description (*r*(7) = .487, *p* = .2, 95% CI [-.261, .870]).

#### c) Item-level correlation between the mean self-perceived accuracy in the six descriptions and the r value of likelihood-base rates for the six descriptions

Using the six correlation coefficients from analysis (b), we calculated Pearson's correlation between these *r* values and the mean self-perceived accuracy reported in the six descriptions. The correlation coefficient for the likelihood-base rates correlations were meant to represent the degree of conformity to base rates when participants made predictions, as mentioned in the target article (Kahneman & Tversky, 1973, p. 240). We found no support for a correlation between self-perceived accuracy and degree of conformity to base rates (*r*(4) = .464, *p* = .4, 95% CI [-.557, .927].

[Placeholder for Bayesian analyses:] To quantify support for the null hypothesis that self-perceived accuracy is not correlated with the degree of conformity to base rates, we conducted a Bayesian analysis and found that the support for the null hypothesis was 1.259 (BF01) times stronger than the alternative hypothesis. (Cauchy prior = .707)

#### d) Item-level correlation between mean base rates and mean likelihood rankings for the null description

The null description was shown to all participants in the likelihood prediction condition. We calculated Pearson's correlation between mean base rates and likelihood rankings across the nine fields of specialization, finding no support for an association between the two (*r*(7) = -.370, *p* = .3, 95% CI [-.830, .390]).

#### e) Item-level correlations between the mean similarity ranking and mean base rates for each of the nine fields of specialization

As an additional analysis, we calculated Pearson's correlation between the mean base rates and the mean similarity ranking across the nine fields of specialization at the item level, repeated for each of the six descriptions shown to the participants. We found no support for a correlation between likelihood rankings and base rates in the business administration description (*r*(7) = .179, *p* = .6, 95% CI [-.551, .753]), computer science description (*r*(7) = -.320, *p* = .4, 95% CI [-.812, .437]), engineering description (*r*(7) = .504, *p* = .2, 95% CI [-.240, .875]), law description (*r*(7) = .195, *p* = .6, 95% CI [-.539, .761]), library science description (*r*(7) = -.410, *p* = .3, 95% CI [-.844, .349]), and medicine description (*r*(7) = .280, *p* = .5, 95% CI [-.472, .796]).

# Study 3: Prior vs individuating evidence

## Design

Study 3 had a between-subjects design between high and low engineer conditions, summarized in Table 10.

###### Table 10 *Study 3 (prior vs individuating evidence): Experimental design [Between subject]*

|  |  |
| --- | --- |
| **High engineer condition**  Participants were given one out of six (including a null condition; even random assignment) descriptions of individuals drawn from a hypothetical sample of 70 engineers and 30 lawyers. | **Low engineer condition**  Participants were given one out of six (including a null condition; even random assignment) descriptions of individuals drawn from a hypothetical sample of 30 engineers and 70 lawyers. |
| Dependent variable:  **Probability prediction**  Participants predicted the probability of the person being an engineer for the provided description.  Scale: 0% to 100% | |

###### 

## Manipulations

In Study 3, participants were evenly and randomly assigned into either the high engineer or low engineer conditions. Participants were told to assume that a panel of psychologists had administered personality tests to 70 (30 for low engineer condition) engineers and 30 (70 for low engineer condition) lawyers and wrote descriptions for all of them. Within the high and low engineer conditions, participants were evenly and randomly split to be shown one out of six descriptions (two for engineers, two for lawyers, one for control with irrelevant information, one for null with no information).

## Data analysis strategy

We adjusted Study 3 so that each participant only rated one profile, whereas in the target each participant rated all profiles in the assigned condition. Given their design, they seemed to have aggregated the profiles and conducted an independent samples t-test comparing the mean judged probabilities in the low versus high engineer condition, not taking into account the different profiles. Given the changed design, and to allow for more nuanced findings that would also show differences between the profiles, we also changed the analyses. We conducted a participant-level 2 (condition) x 6 (description) two-way between subject ANOVA analysis with judged probability as the dependent variable, and post-hoc Tukey’s HSD tests.

We also added a complementary item-level paired samples t-test comparing the mean judged probabilities combined from the five descriptions (except the null) between the high and low engineer condition. We note that given the small number of profiles, the effect would have to be very large to be detectable (*d* ~> 1.3), and so we consider the two-way ANOVA participant-level analysis the closer and accurate test.

Finally, we mirrored the target article’s reporting in our descriptive analysis scatter plot of median judged probabilities in high against low engineer condition for descriptive analysis against the Bayesian prediction curve.

## Results and discussion

[Reminder for Stage 2: Alpha is set to .005]

We summarized the descriptives of the judged probabilities of each description in the low and high engineer conditions in Table 11.

###### Table 11 *Study 3: Descriptives of judged probabilities in low and high engineer conditions*

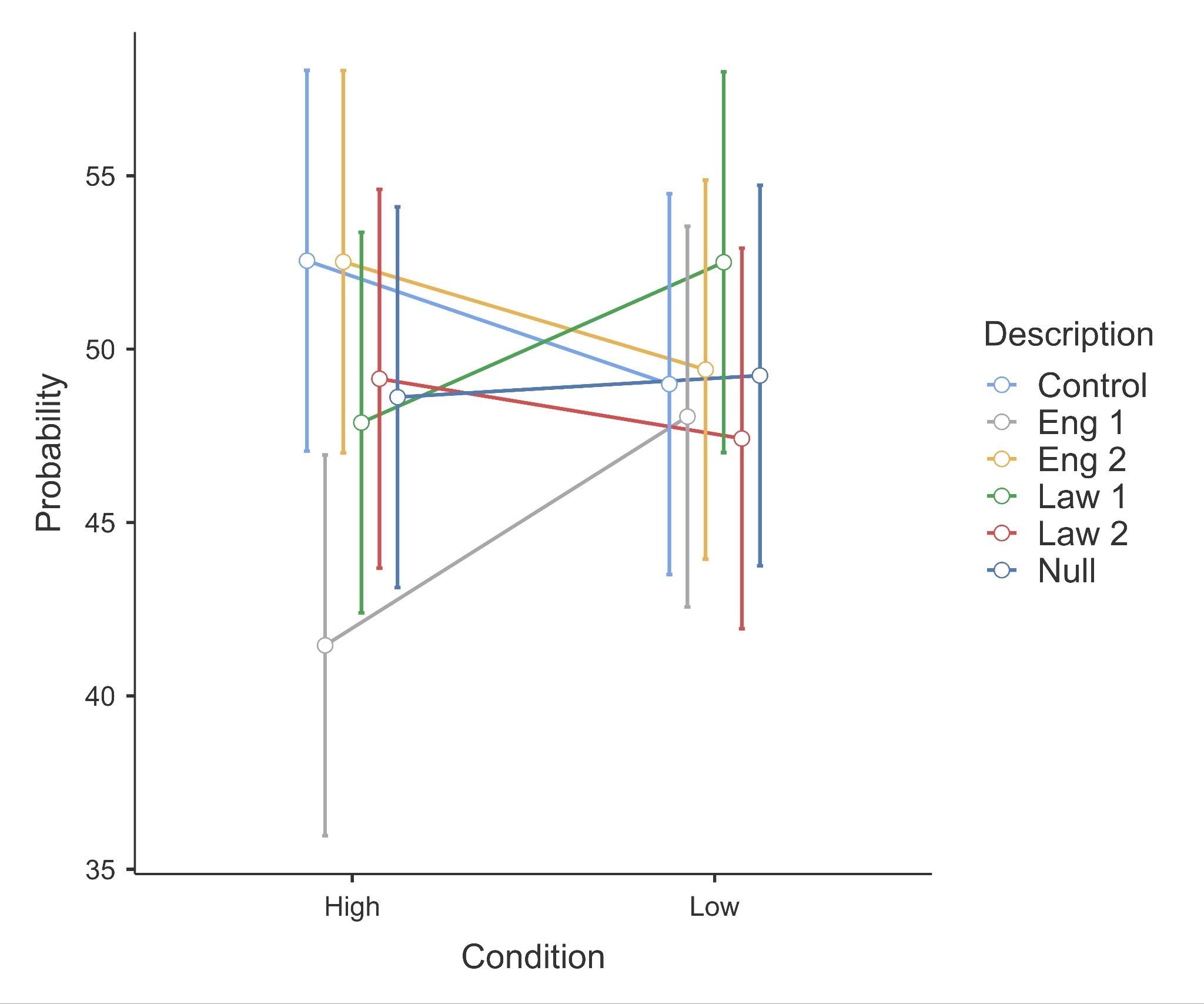
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Low engineer (*n* = 654)** | | | **High engineer (*n* = 655)** | | |
|  | *n* | *M* (%) | *SD* (%) | *n* | *M* (%) | *SD* (%) |
| Engineer 1 | 109 | 50.248 | 29.58 | 109 | 47.835 | 28.01 |
| Engineer 2 | 109 | 47.532 | 29.68 | 109 | 56.312 | 29.37 |
| Lawyer 1 | 109 | 50.119 | 29.39 | 109 | 50.257 | 28.72 |
| Lawyer 2 | 109 | 49.734 | 29.71 | 109 | 47.193 | 29.32 |
| Control | 109 | 49.064 | 28.44 | 110 | 55.009 | 30.92 |
| Null | 109 | 52.743 | 29.09 | 109 | 53.266 | 28.79 |

We conducted a 2 (high versus low) x 6 (descriptions) two-way between ANOVA using judged probabilities as the dependent variable (see Figure 1). We found no support for the main effect of condition (high or low engineer) (*F*(1, 1297) = 1.155, *p* = .282, η2 = .001, ⍵2 = .000), the main effect of description (*F*(5, 1297) = 0.844, *p* = .518, η2 = .003, ⍵2 = -.001), and the interaction effect (*F*(5, 1297) = 1.361, *p* = .236, η2 = .005, ⍵2 = .001). Since no main or interaction effects were supported, post-hoc tests were not conducted.

[Note for Stage 2: However, if the representativeness hypothesis is true, we expect there to be no main effect of condition but the main effect of description would be supported. Furthermore, to investigate the target article’s claim that worthless evidence leads to the ignoring of prior evidence whilst a complete lack of information does not (Kahneman & Tversky, 1973, p. 242), we plan to conduct post-hoc tests comparing the mean judged probabilities in the null versus the control descriptions].

[Placeholder for Bayesian analyses:] To quantify support for the null effect that the condition (high or low engineer) did not affect participants’ judgements of probability, a Bayesian analysis was conducted for the 2 (high versus low) x 6 (descriptions) two-way between ANOVA using JAMOVI. We found that the support for the null hypothesis was 9.09 times larger than the alternative hypothesis that there exists differences in judged probabilities between the high and low engineer conditions.

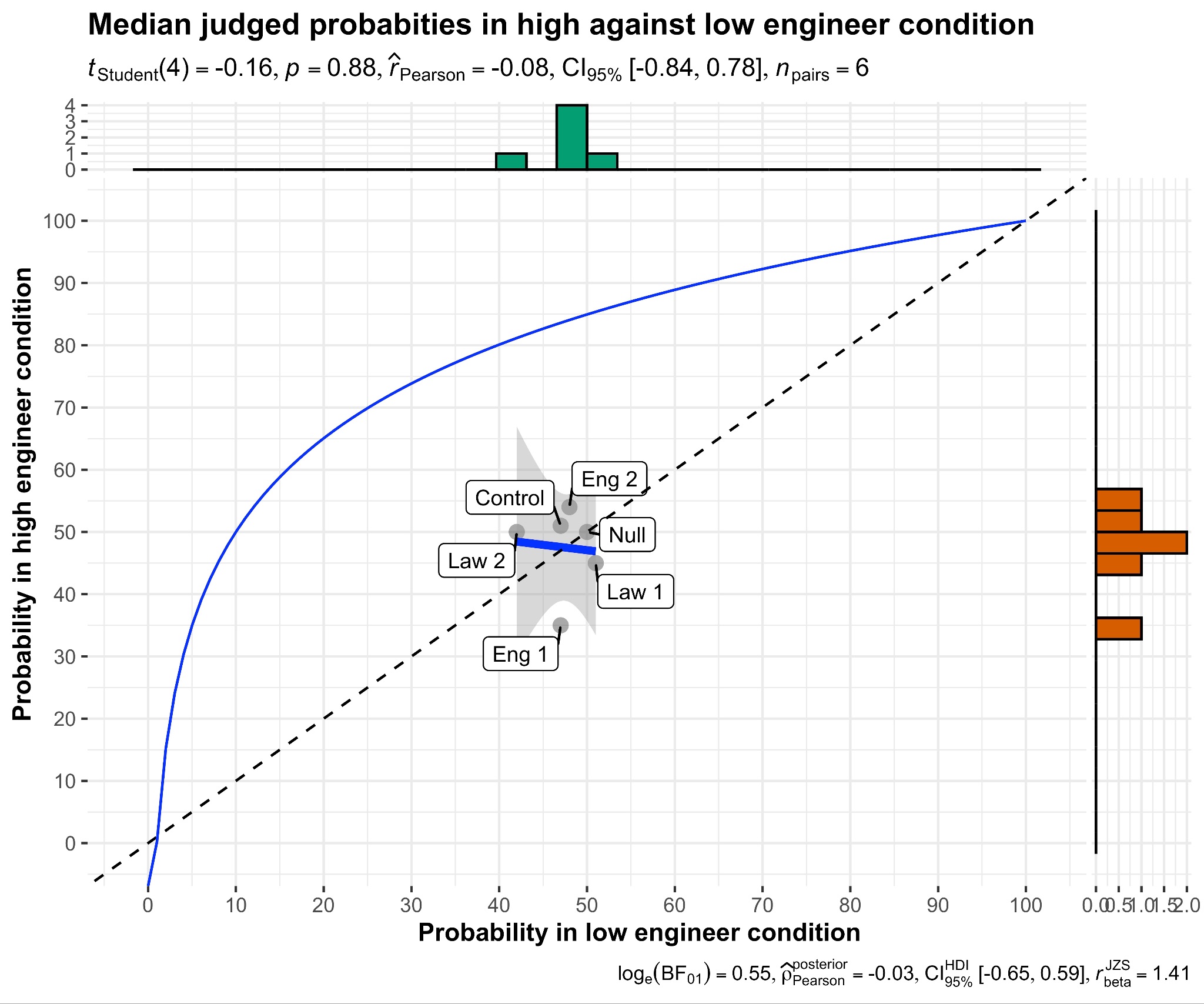
###### Figure 1 *Study 3: Interaction between base rates (high or low engineer) and descriptions on probability*



*Note*. Created with JAMOVI [2.4.12.0] (The JAMOVI project, 2023).

We mirrored Figure 2 after the plot provided in the target article, showing the scatter plot of median judged probabilities in high against low engineer condition for descriptive analysis against the curved line of the correct relation according to Bayes’ rule. The data points did not follow the identity line nor the Bayesian line, which is in contrast to the target article which reported that all data points (except the null) conformed to the identity line. Moreover, the point for the null condition did not fall on the 70/30 point on the Bayesian as described in the target article.

###### Figure 2 *Study 3: Scatter plot of median judged probabilities in high against low condition*



*Note*. Created with ggstatsplot package (Patil, 2021).

We supplemented the participant-level analysis with an item-level analysis. We calculated the mean judged probabilities for each description in the low engineer condition (e.g., engineer 1, lawyer 1, etc.), repeated for the high engineer condition, resulting in five pairs (excluding the null) of mean judged probabilities. We conducted an item-level paired samples t-test and found no support that the mean judged probabilities in the low engineer (*M* = 49.91%, *SD* = 1.708%) and high engineer (*M* = 51.65%, *SD* = 3.795%) conditions differed (*t*(4) = 0.865, *p* = .436, *d* = 0.387, 95% CI[-0.548, 1.279]).

In comparison to the target article, an independent samples t-test of the aggregate of profiles found support for a mean difference in judged probability between the low engineer (*M* = 50%) and high engineer (*M* = 55%) conditions (*t*(169) = 3.23, *p* < .01).

# Study 4: Prediction vs evaluation

## Design

Study 4 had a 2 (adjectives/reports) x 2 (evaluations/ predictions) design, summarized in Table 12.

###### 

###### Table 12 *Study 4 (prediction vs evaluation): Experimental design [Between subject]*

|  |  |  |
| --- | --- | --- |
| IV1: Prediction vs evaluation [between]  IV2: Adjectives vs reports [between] | IV1: **Evaluation**  Participants evaluated the described student’s academic ability compared to the other students in the entire class. | IV1: **Prediction**  Participants predicted the described students’ class standing compared to the other students in the entire class. |
| IV2: **Adjectives**  Participants were given one out of nine (even random assignment) descriptions of college freshmen each containing five adjectives. | Dependent variables: **Evaluation percentage** [replication]  Scale:  0% = *The student described is more impressive than 0% of the class*;  100% = *The student described is more impressive than 100% of the class*  **Confidence** [extension]  Participants rated their confidence in their estimation.  Scale:  0% = *Not at all confident*;  100% = *Perfectly confident* | Dependent variables: **Prediction percentage** [replication]  Scale:  0% = *The described student’s GPA is the lowest in the class*;  100% = *The described student’s GPA is the highest in the class*.  **Confidence** [extension]  Participants rated their confidence in their estimation.  Scale:  0% = *Not at all confident*;  100% = *Perfectly confident* |
| IV2: **Reports**  Participants were given one out of nine (even random assignment) nine descriptions of college freshmen each containing a paragraph length report. |

## Manipulations

Participants were evenly and randomly assigned into the adjectives-evaluations condition, adjectives-predictions condition, reports-evaluations condition, or the reports-predictions condition.

### Adjectives vs reports

In the adjectives conditions, participants were shown one out of nine descriptions of a college freshman containing five adjectives describing their intellectual quality and character (three positive, three mixed, three negative). In the reports conditions, the descriptions were instead paragraph-length reports based on the adjectives description, yet with additional information about the student’s background and adjustment to college.

### Evaluation vs prediction

In the evaluation conditions, participants evaluated how impressive the described student was to them in terms of academic ability compared to the other students in the entire class. In the prediction conditions, participants were asked to predict the student's class standing in terms of grade point average as a percentile.

## Measures

### Replication

#### Evaluation percentage

We adjusted the scales and wordings used in the target article. The original phrasing instructed participants to estimate “the percentage of students in the entire class whose descriptions indicate a higher academic ability.” (Kahneman & Tversky, 1973, p. 244). However, this seemed to be counter-intuitive in direction compared to the plots of the results (p. 244), with a lower percentile score implying being more impressed. Therefore, we modified it to have participants “estimate, in terms of a percentile score, how impressive the described student is to you in terms of academic ability compared to the other students in the entire class (Example: a percentile score of 65 means that you consider this student to be more impressive in terms of academic ability than 65% of his class)”. This aimed to clarify that a higher percentage corresponds to a more impressive evaluation. (0% = *The student described is more impressive than 0% of the class*; 100% = *The student described is more impressive than 100% of the class*).

#### Prediction percentage

Participants predicted the described students’ class standing in terms of his grade point average in percentiles (0% = *The described student’s GPA is the lowest in the class*; 100% = *The described student’s GPA is the highest in the class*).

### Extension

#### Confidence

We added a measure of participants’ self-reported confidence in their predictions to directly measure the concept of certainty in predictions compared to the indirect approach of inferring through SD in predictions in the target article. Participants self-reported their confidence in their estimations (0% = *Not at all confident*; 100% = *Perfectly confident*).

## Data analysis strategy

### Replication

We adjusted the analyses in the target article which ran t-tests comparing SD between the prediction and evaluation conditions due to practical constraints with our study design. In order to address the same research question of differences in SD between the conditions, we ran a) Levene's test of homogeneity to test for differences in SD between the prediction and evaluation conditions in the 2 (adjectives/ reports) x 2 (evaluation/ prediction) design. We mirrored the target article with b) A descriptive analysis of plotting the mean prediction against evaluation in the adjectives and reports conditions.

### Extension

We extended the target article to run a 2 (adjective/report) x 2 (evaluation/prediction) ANOVA with confidence in prediction as the dependent variable. This was to address the issue in the target article where the concept of uncertainty in predictions was measured through indirectly inferring from differences in SD of predictions. We aimed to directly ask participants to report their confidence and analyze the differences between the four conditions.

## Results and discussion

[Reminder for Stage 2: Alpha is set to .005]

### Replication

We summarized the mean evaluations and predictions in the adjectives and reports conditions for all descriptions in Table 13.

###### 

###### Table 13 *Study 4: Mean evaluations and predictions in the four conditions*

|  | **Adjectives-evaluation (*n* = 327)** | | **Adjectives-prediction**  **(*n* = 328)** | | **Reports-evaluation  (*n* = 327)** | | **Reports-prediction**  **(*n* = 327)** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *n* | *M* (%) | *n* | *M* (%) | *n* | *M* (%) | *n* | *M* (%) |
| Positive original | 36 | 51.03 | 37 | 54.89 | 37 | 56.68 | 37 | 55.49 |
| Positive 1 | 36 | 50.11 | 36 | 51.31 | 36 | 46.50 | 36 | 52.97 |
| Positive 2 | 37 | 35.81 | 37 | 43.60 | 36 | 50.69 | 36 | 55.33 |
| Mixed 1 | 37 | 56.16 | 37 | 49.54 | 37 | 52.38 | 37 | 45.11 |
| Mixed 2 | 36 | 42.08 | 37 | 48.65 | 36 | 51.11 | 36 | 48.86 |
| Mixed 3 | 37 | 49.68 | 36 | 53.97 | 37 | 44.76 | 37 | 58.08 |
| Negative 1 | 36 | 53.00 | 36 | 44.03 | 36 | 50.86 | 36 | 48.56 |
| Negative 2 | 36 | 64.11 | 36 | 47.83 | 36 | 50.89 | 36 | 51.14 |
| Negative 3 | 36 | 57.19 | 36 | 43.56 | 36 | 56.44 | 36 | 45.97 |

#### a) Differences in SD between the prediction and evaluation conditions

We conducted the Levene’s test of homogeneity to test for differences in SD between the prediction and evaluation conditions in the 2 (adjectives/ reports) x 2 (evaluation/ prediction) design. At the adjectives level, the test of homogeneity indicated equal variances between the evaluation and prediction conditions (*F*(1, 653) = 0.97, *p* = .32). At the reports level, the test of homogeneity also indicated equal variances between the evaluation and prediction conditions (*F*(1, 652) = 0.05, *p* = .82).

[Placeholder for Bayesian analyses:] To quantify support for the null hypothesis of equal variances, we converted the *F* value from the Levene’s tests of homogeneity into a Bayes factor (BF01). At the adjectives level, the observed data favored the null hypothesis compared to the alternative hypothesis by a factor of 7.13 for the comparison of variances between the prediction and evaluation conditions. At the reports level, the observed data favored the null hypothesis compared to the alternative hypothesis by a factor of 11.17.

#### b) Plotting mean prediction against evaluation in the adjectives and reports conditions

We mirrored the target article and plotted the scatter plots of mean predicted GPA in the prediction condition against mean evaluated percentile score for the adjective (Figure 3) and report conditions (Figure 4). It can be observed from the two plots that the data points did not conform to the identity line, contrary to the descriptive results reported in the target article.

###### Figure 3 *Study 4: Mean predicted GPA against mean evaluation for adjectives conditio*n

*Note*. Created with ggstatsplot package (Patil, 2021). GPA: 0% = *The described student’s GPA is the lowest in the class*; 100% = *The described student’s GPA is the highest in the class*; Evaluation: 0% = *The student described is more impressive than 0% of the class*; 100% = *The student described is more impressive than 100% of the class*.

###### Figure 4 *Study 4: Mean predicted GPA against mean evaluation for reports conditio*n

*Note*. Created with ggstatsplot package (Patil, 2021). GPA: 0% = *The described student’s GPA is the lowest in the class*; 100% = *The described student’s GPA is the highest in the class*; Evaluation: 0% = *The student described is more impressive than 0% of the class*; 100% = *The student described is more impressive than 100% of the class*.

### 

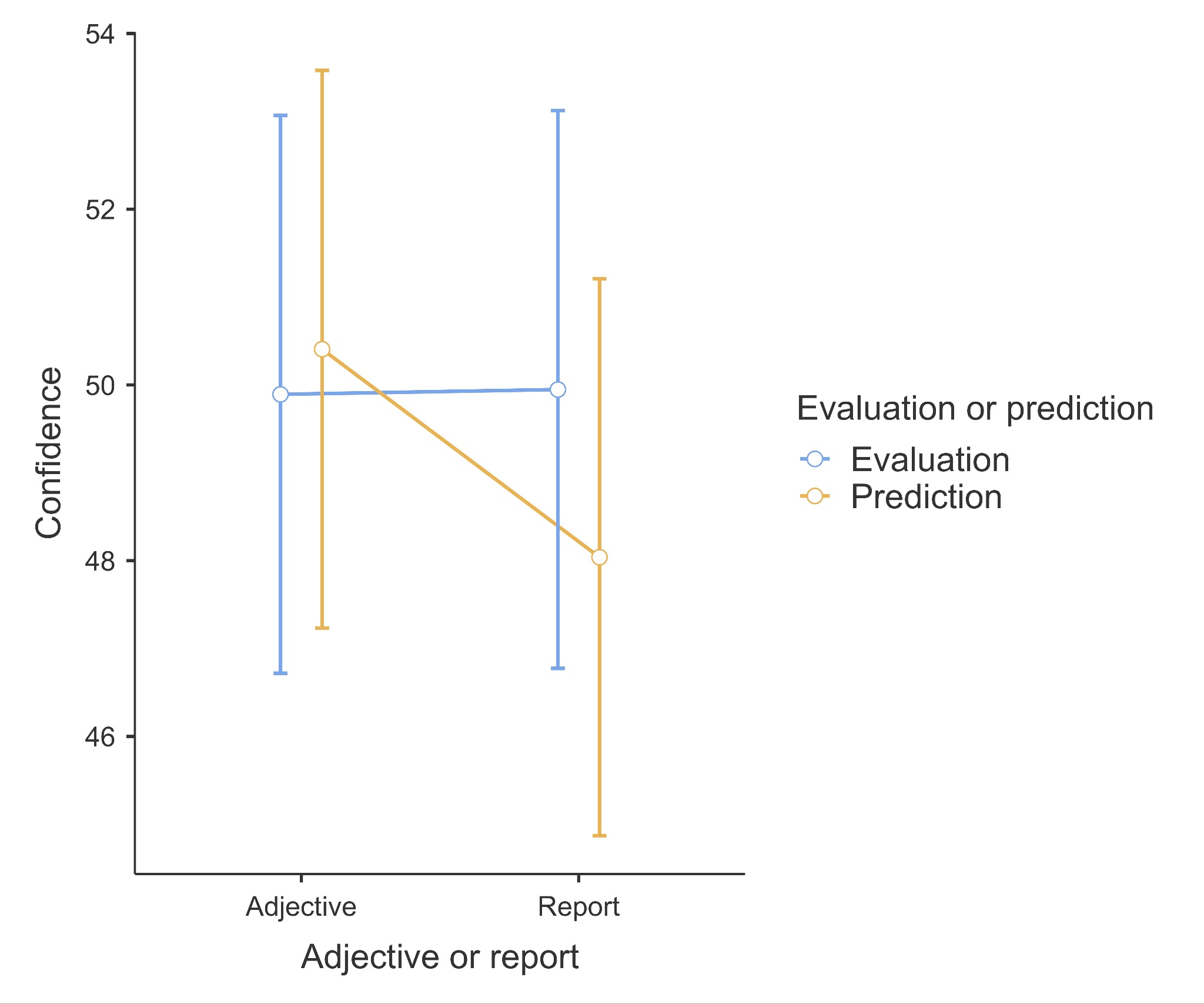
### Extension

We extended the target article to run a 2 (adjective/report) x 2 (evaluation/prediction) two-way between ANOVA with confidence in prediction as the dependent variable (Figure 5). We failed to find support for the main effect of the adjective and report manipulation on confidence in predictions (*F*(1, 1305) = 0.413, *p* = .521, η2 = .000, ⍵2 = -.000). We also failed to find support for the main effect of the evaluation and prediction manipulation on confidence in predictions (*F*(1, 1305) = 0.046, *p* =.150, η2 = .000, ⍵2 = -.001). There was also no support for the interaction effect (*F*(1, 1305) = 2.073, *p* = .150, η2 = .002, ⍵2 = .001).

[Note for Stage 2: If the representativeness hypothesis is true, we expect this to be a null effect. In that case, we will see if this extension result agrees with the results from the Levene’s test in analysis a, then run Bayesian analyses to investigate the null effect].

[Placeholder for Bayesian analyses:] To quantify support for the null hypothesis that confidence was not affected by whether an evaluation or a prediction was being made, we ran a Bayesian analysis for the 2 (adjective/report) x 2 (evaluation/prediction) two-way between ANOVA. It was found that the support for the null hypothesis was 14.7 times stronger than the alternative hypothesis.

###### Figure 5 *Study 4: Adjective, reports, evaluation, and prediction: Impact on mean confidence*



*Note*. Created with JAMOVI [2.4.12.0] (The JAMOVI project, 2023). Confidence: 0% = *Not at all confident*; 100% = *Perfectly confident*

# 

# Study 5: Prediction vs translation

## Design

Study 5 had a between-subjects design contrasting between three conditions (even and random assignment): academic achievement, mental concentration, and sense of humor, summarized in Table 14.

###### Table 14 *Study 5 (prediction vs translation): Experimental design [Between subject]*

|  |  |  |
| --- | --- | --- |
| **Academic achievement**  Participants were given 10 percentile scores representing the academic achievements of different freshmen in their first year. | **Mental concentration**  Participants were given 10 percentile scores representing the scores on a mental concentration test of different freshmen in their first year and told that the test is unreliable. | **Sense of humor**  Participants were given 10 percentile scores representing the sense of humor of different freshmen in their first year and told that sense of humor is not an accurate predictor of GPA. |
| Dependent variable: **Predicted GPA**  Participants were told to give their best guesses of each of the 10 students’ GPA in their second year.  Scale:  0.0 = *The student scored the lowest possible grade for all courses in the second year*;  4.0 = *The student scored the highest possible grade for all courses in the second year*. | | |

## Measures

We made slight adjustments to Study 5. We included an explanation of GPA in case participants were not familiar with the concept, and we deviated from the original prompt of “... give your best guess about his grade point average for *that* year” (Kahneman and Tversky, 1973, p. 245), to “... give your best guess about his grade point average for the *second* year” (the changed word is italicized) (0.0 to 4.0). This was to address the ambiguity and possible confusion regarding the focal year.

## Data analysis strategy

We mirrored the target article and conducted a) t-tests comparing mean predicted GPA between the academic achievement and mental concentration condition, and that between the mental concentration and sense of humor condition; b) t-tests comparing SD between the academic achievement and mental concentration condition, and that between the mental concentration and sense of humor condition; c) t-tests comparing correlations between the academic achievement and mental concentration condition, and that between the mental concentration and sense of humor condition; and d) descriptive analysis: plot of mean predicted GPA against given percentile scores in the three conditions.

## Results and discussion

[Reminder for Stage 2: Alpha is set to .005]

We summarized the descriptives for the within-participants means, SD, and correlations with the input percentile scores for predicted GPA in Table 15, and the findings in Table 17.

###### 

###### Table 15 *Study 5: Means and SDs of within-participants means, SD, and correlation with input scores*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Academic achievement**  **(*n* = 436)** | | **Mental concentration**  **(*n* = 437)** | | **Sense of humor**  **(*n* = 436)** | |
|  | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Mean predicted GPA | 1.97 | 0.45 | 1.99 | 0.44 | 2.02 | 0.44 |
| *SD* | 1.39 | 0.22 | 1.39 | 0.21 | 1.40 | 0.23 |
| *r* | 0.00 | 0.33 | 0.01 | 0.34 | 0.00 | 0.35 |

*Note*. *M* = between participants means. *SD* = between participants SD.

#### a) Mean differences in predicted GPA

Mirroring the target article, we conducted independent sample t-tests to compare the within participants mean predicted GPA in the academic achievement condition (*M* = 1.97, *SD* = 0.45) against that in the mental concentration condition (*M* = 1.99, *SD* = 0.44), finding no support for a difference (*t*(871) = -0.79, *p* = .4, *d* = -0.05, 95% CI [-0.19, 0.08]). We also found no support for a difference in mean predicted GPA between the mental concentration condition and the sense of humor condition (*M* = 2.02, *SD* = 0.44) (*t*(871) = -1.2, *p* = .2, *d* = -0.08, 95% CI [-0.22, 0.05]).

[Placeholder for Bayesian analyses:] To quantify support for the null effect that within participants means do not differ between the academic achievement and mental concentration condition, we conducted a Bayesian independent samples t-test. We found that the support for the null hypothesis was 9.74 times stronger than that for the alternative hypothesis (Cauchy prior = .707).

#### b) Mean differences in SD of predicted GPA

Mirroring the target article, we conducted independent sample t-tests to compare the within participants SD of predicted GPA in the academic achievement condition (*M* = 1.39, *SD* = 0.22) against that in the mental concentration condition (*M* = 1.39, *SD* = 0.21), finding no support for a difference (*t*(871) = 0.22, *p* = .8, *d* = 0.01, 95% CI [-0.12, 0.15]). We also found no support for a difference in mean predicted GPA between the mental concentration condition and the sense of humor condition (*M* = 1.40, *SD* = 0.23) (*t*(871) = -0.9, *p* = .4, *d* = -0.06, 95% CI [-0.19, 0.07]).

[Placeholder for Bayesian analyses:] To quantify support for the null effect that within participants SD do not differ between the academic achievement and mental concentration condition, we conducted a Bayesian independent samples t-test. We found that the support for the null hypothesis was 12.91 times stronger than that for the alternative hypothesis (Cauchy prior = .707).

#### c) Mean differences in correlations of predicted GPA with input scores

Mirroring the target article, we conducted an independent samples t-tests to compare the within participants correlation between predicted GPA and the input scores in the academic achievement condition (*M* = .00, *SD* = .33) against that in the mental concentration condition (*M* = .01, *SD* = .34), finding no support for a difference (*t*(871) = -0.17, *p* = .9, *d* = -0.01, 95% CI [-0.14, 0.12]). We also found no support for a difference in mean predicted GPA between the mental concentration condition and the sense of humor condition (*M* = -.00, *SD* = .35) (*t*(871) = 0.26, *p* = .8, *d* = 0.02, 95% CI [-0.12, 0.15]).

[Placeholder for Bayesian analyses:] To quantify support for the null effect that within participants correlations do not differ between the academic achievement and mental concentration condition, we conducted a Bayesian independent samples t-test. We found that the support for the null hypothesis was 13.03 times stronger than that for the alternative hypothesis (Cauchy prior = .707).

#### d) Plot of mean predicted GPA against given percentile scores in the three conditions

We mirrored the plot in the target article of mean predicted GPA against the 10 given percentile scores in the three conditions as a descriptive analysis (Figure 6). The values of the mean predicted GPA in the three conditions and the percentile scores used are shown in Table 16. In the target article, based on the descriptive plot, it was claimed that the predictions in the mental concentration condition were not more regressive than that in the academic achievement condition, in contrast to the regressive prediction patterns of the sense of humor condition. However, there did not seem to be differences in the pattern of predictions observable from the replicated plot [due to randomly generated data from Qualtrics].

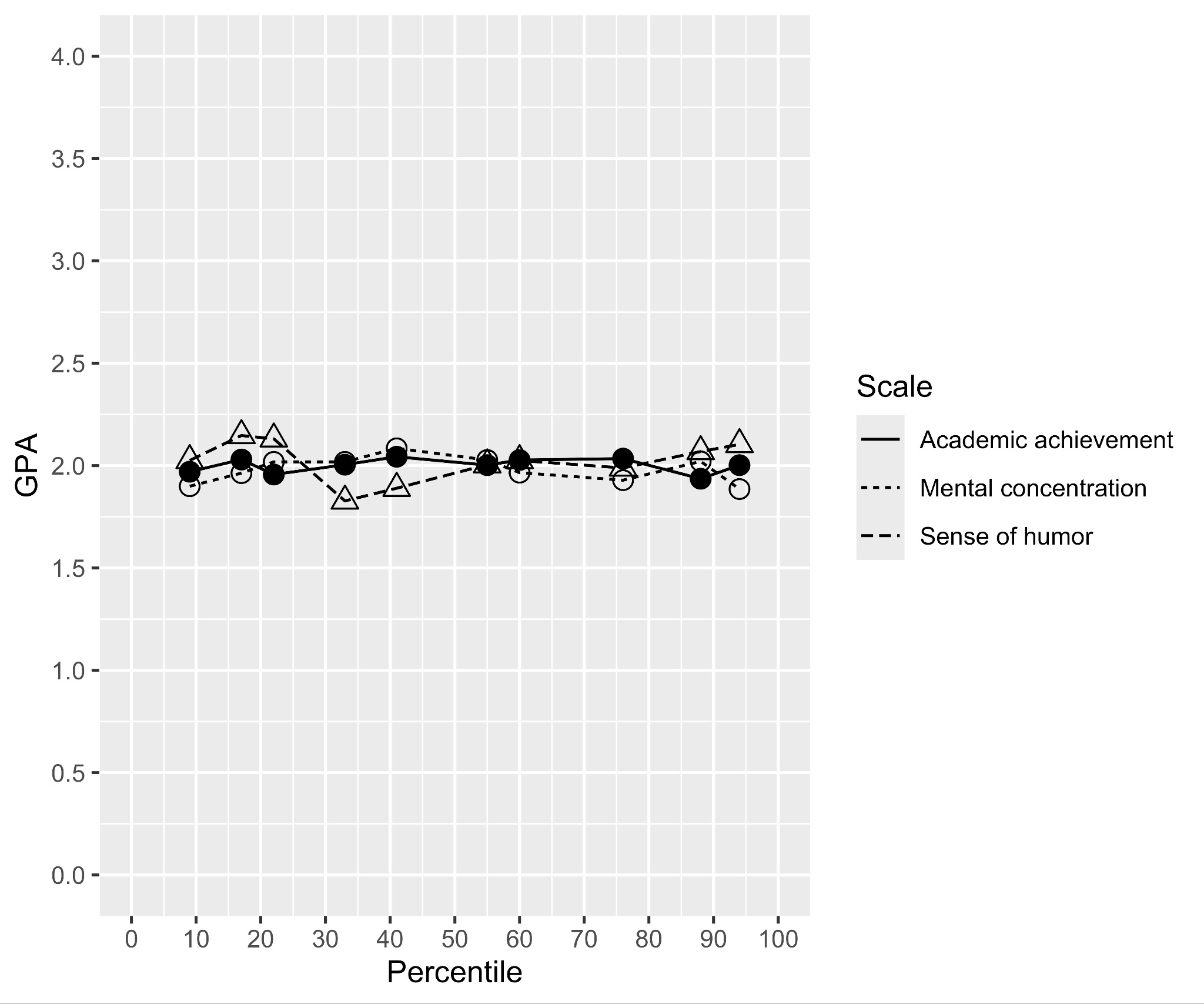
###### 

###### Table 16 *Study 5: Mean predicted GPA in all three conditions*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Student** | **Percentile** | **Academic achievement** | **Mental concentration** | **Sense of humor** |
| A | 9 | 1.90 | 2.00 | 2.03 |
| B | 17 | 1.87 | 1.97 | 1.98 |
| C | 22 | 2.01 | 1.99 | 2.07 |
| D | 33 | 2.06 | 1.93 | 2.02 |
| E | 41 | 2.06 | 2.10 | 1.93 |
| F | 55 | 1.91 | 1.92 | 2.12 |
| G | 60 | 1.96 | 2.01 | 2.05 |
| H | 76 | 1.96 | 1.92 | 2.00 |
| I | 88 | 2.06 | 2.02 | 2.04 |
| J | 94 | 1.85 | 2.03 | 2.03 |

Note. Student = the label of the hypothetical student corresponding to the percentile score shown to participants.

###### Figure 6 *Study 5: Mean predicted GPA against given percentile scores in all three conditions*



*Note*. Created with ggplot2 R [4.3.2] (R Core Team, 2022) package (Wickham, 2016).

###### Table 17 *Study 5: Summary of statistical tests and effects*

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Replication** | | | | | | | | | **Target article** |  |
| **Analysis** | **Condition 1** | | | **Condition 2** | | | ***t*** | ***p*** | ***d* and 95% *CI*** | **Effect** | **Interpretation** |
|  | Name | *M* | *SD* | Name | *M* | *SD* |  |  |  |  |  |
| **a) Mean predicted GPA** | Academic achievement | 1.97 | 0.45 | Mental concentration | 1.99 | 0.44 | -0.79 | .4 | -0.05[-0.19, 0.08] | null effect | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] |
|  | Mental concentration | 1.99 | 0.44 | Sense of humor | 2.03 | 0.44 | 1.2 | .2 | -0.08[-0.22, 0.05] | humor > mental\* | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] |
| **b) SD predicted GPA** | Academic achievement | 1.39 | 0.22 | Mental concentration | 1.39 | 0.21 | 0.22 | .8 | 0.01[-0.12, 0.15] | null effect | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] |
|  | Mental concentration | 1.39 | 0.21 | Sense of humor | 1.40 | 0.23 | -0.9 | .4 | -0.06[-0.19, 0.07] | humor < mental | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] |
| **c) r with predicted GPA** | Academic achievement | 0.00 | 0.33 | Mental concentration | 0.01 | 0.34 | -0.17 | .9 | -0.01[-0.14, 0.12] | null effect | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] |
|  | Mental concentration | 0.01 | 0.34 | Sense of humor | 0.00 | 0.35 | 0.26 | .8 | 0.02[-0.12, 0.15] | effect not supported | N/A |

*Note*. Independent samples t-test, *df* = 871. *CI* = 95% confidence intervals. The interpretation of outcome was based on LeBel et al. (2019). \* “Humor > mental” = the original authors found support that the mean in the sense of humor condition was greater than the mental concentration condition, but the effect was not specified.

# Study 6: Effect of consistency

## Design

The design of Study 6 was summarized in Table 18.

###### Table 18 *Study 6 (effect of consistency): Experimental design [within-subject]*

|  |  |  |
| --- | --- | --- |
| **Correlated aptitude tests**  IV group participants were given a student’s scores on two aptitude tests supposedly correlated with each other. | **Uncorrelated aptitude tests**  The same IV group participants were given a student’s scores on two other aptitude tests supposedly uncorrelated with each other. | **Control**  A separate control group of participants were given the same two pairs of tests but the pair that was told to be correlated was switched to being uncorrelated and vice versa. |
| Dependent variables:  **Predicted GPA**  Participants predicted the students’ GPA.  Scale: 0.0 = *The student scored the lowest possible grade for all courses*; 4.0 = *The student scored the highest possible grade for all courses*  **Confidence**  Participants rated their confidence in their estimation.  Scale:  0% = *Not at all confident;*  100% = *Perfectly confident* | | |
|

## 

## Manipulations

Participants were evenly and randomly assigned into either the experimental or control condition. In the experimental condition, participants were given a college student’s scores on one pair of aptitude tests (creative thinking and symbolic ability) assumed to be correlated, and another college student’s scores on another pair of aptitude tests (mental flexibility and systematic reasoning) assumed to be uncorrelated. They were told to assume that all these tests are equally successful in predicting college performance. In the control condition, all procedures and materials were identical except that the labels for the pairs of aptitude tests were switched (i.e., creative and symbolic ability became uncorrelated, and mental flexibility and systematic reasoning was correlated).

## Measures

### GPA

The target article did not specify the scale used, and so we reverted to using the same scale for GPA from Study 5. Participants predicted the 10 described students’ GPA (0.0 to 4.0).

### Confidence

The target article did not specify the scale used. We opted to use a simple percentage scale since it was a relatively straightforward way of self-reporting confidence that the average participant should be familiar with (0% = *Not at all confident*; 100% = *Perfectly confident*).

## Data analysis strategy

We mirrored the analyses in the target article, conducting a t-test comparing the confidence in prediction in the correlated aptitude tests versus the uncorrelated aptitude tests conditions (repeated analyses again with the control group).

## 

## Results and discussion

[Reminder for Stage 2: Alpha is set to .005]

We summarized the descriptives for participants’ confidence in prediction in the experimental and control condition in Table 19.

###### Table 19 *Study 6: Descriptives for confidence in predictions in the experimental and control condition*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Experimental (*n* = 655)** | | **Control (*n* = 654)** | |
|  | **Correlated** | **Uncorrelated** | **Correlated** | **Uncorrelated** |
| *M* | 49.25 | 49.67 | 49.98 | 47.96 |
| *SD* | 28.55 | 29.23 | 28.53 | 29.21 |

We conducted a paired sample t-test comparing confidence in predictions in the experimental condition when predicting with correlated (*M* = 49.25, *SD* = 28.55) versus uncorrelated (*M* = 49.67, *SD* = 29.23) pairs of aptitude tests, finding no support for a difference (*t*(654) = -0.269, *p* =.788, *d* = -0.011, 95% CI [-0.087, 0.066]). We repeated the analyses in the control condition where the labels of the correlated and uncorrelated tests were switched, finding no support for a difference in confidence in predictions when predicting from correlated (*M* = 49.98, *SD* = 28.53) versus uncorrelated (*M* = 47.96, *SD* = 29.21) pairs of aptitude tests (*t*(653) = -1.284, *p* = .200, *d* = -0.050, 95% CI [-0.127, 0.027].

# Study 7: Persistence of non-regressive intuitions

## Design

Study 7 had a correlational design where all participants were measured for degree of regression and level of statistical knowledge, usage, and training, as summarized in Table 20.

###### Table 20 *Study 7 (persistence of non-regressive intuitions): Experimental design*

|  |
| --- |
| IV1: **Statistical knowledge** [extension]  Participants rated their level of statistical knowledge.  Scale: 0 = *Not at all proficient in statistics*; 100 = *Very proficient in statistics*.  IV2: **Statistics usage** [extension]  Participants rated their frequency of using statistics in their job/life.  Scale: 0 = *Not at all*; 100 = *All the time*.  IV3: **Statistics training** [extension]  Participants reported their level of statistics training.  Scale: *0 = No statistics training; 1 = High-school level statistics training; 2 = College level statistics training; 3 = Professional training in statistics; 4 = Academic training in statistics (postgraduate and above)*  DV: **Degree of regression** [replication]  Participants were asked to provide a 95% confidence interval for a score of 140 on an IQ test.  The lowest bound of the interval is 0, and the upper bound is unlimited.  The average of the two bounds will be taken to estimate the degree of regression. |

## Data analysis strategy

### Replication

We mirrored the analyses using descriptives reported by the original authors and calculated the proportions of participants stating 95% CI that were symmetric around 140, regressive, or counter-regressive.

### Extension

We extended the target article and calculated three separate correlations between degree of regression and self-reported level of statistical knowledge, usage, and training. This was to address the issue in the target article’s study design that it may have been vulnerable to artifacts (e.g., poor learning of statistical knowledge in the sample of psychology students). We focused on self-reported level of statistical knowledge as the main criterion for determining support for the alternative hypothesis that statistics proficiency correlates with degree of regression.

## Results and discussion

[Reminder for Stage 2: Alpha is set to .005]

### Replication

We mirrored the original study and calculated the proportions of participants that stated confidence intervals that were symmetric around 140, regressive, or counter-regressive. The original authors did not specify what constituted a regressive or counter-regressive 95% CI. Therefore, we averaged the lower and upper bounds of the 95% CI provided by the participants to obtain a point of symmetry for each prediction, and considered those with a point of symmetry lower than 140 to be regressive 95% CIs, and those above 140 to be counter-regressive 95% CIs. Among the 1309 participants, we found that 0% provided symmetric 95% CIs, 0% provided regressive 95% CI, and 100% provided counter-regressive 95% CI.

[Note for reviewers: data was skewed due to unrealistically large values generated by Qualtrics’ simulation with the question validation disabled.].

Although the original authors expected that most participants would fail to make regressive estimates, implying that higher degree of regression was desirable, we considered the possibility that some participants may also fail to grasp the concept of regression by making overly regressive estimates. Therefore, we repeated the analyses but excluded overly regressive estimates (i.e., 95% CI with a point of symmetry lower than 100), finding that out of 1309 participants, 0% provided symmetric 95% CIs, 0% provided regressive 95% CI, and 100% provided counter-regressive 95% CI.

### Extension

We extended the target article and calculated the correlations between degree of regression and self-reported level of statistical knowledge. We obtained the values for degree of regression by finding the point of symmetry of participants’ provided 95% CI, then subtracting this from 140. Therefore, a higher value corresponded to a more regressive estimate. A Pearson’s correlation analysis found no support that degree of regression was associated with self-reported level of statistical knowledge (*r*(1307) = -.009, *p* = .733, 95% CI [-.063, .045]) (Figure 7). We repeated the analysis excluding overly regressive estimates (i.e., 95% CIs with a point of symmetry lower than 100), finding no support for degree of regression as associated with self-reported level of statistical knowledge (*r*(1307) = -.009, *p* = .733, 95% CI [-.063, .045]) (Figure 8). [Placeholder for Bayesian analyses:] To quantify support for the null hypothesis that statistical knowledge is not related to degree of regression, we conducted a Bayesian analysis and found that support for the null hypothesis (BF01) was 0.138 times stronger when all data was included, and the factor was 0.138 times in favor of the null when overly regressive estimates were excluded (Cauchy prior = .707).

###### Figure 7 *Study 7: Correlation between statistical knowledge and degree of regression*

*Note*. Created with ggstatsplot package (Patil, 2021).

###### Figure 8 *Study 7: Correlation between statistical knowledge and degree of regression (excluding overly regressive estimates)*

###### 

*Note*. Created with ggstatsplot package (Patil, 2021).

For statistics usage, a Pearson’s correlation analysis found no support that degree of regression was associated with self-reported level of statistics usage (*r*(1307) = -.009, *p* = .733, 95% CI [-.063, .045]). We repeated the analysis excluding overly regressive estimates, finding no support for degree of regression as associated with self-reported level of statistics usage (*r*(1307) = -.009, *p* = .733, 95% CI [-.063, .045]).

For statistics training, a Pearson’s correlation analysis found no support that degree of regression was associated with self-reported level of statistics usage (*r*(1307) = -.009, *p* = .733, 95% CI [-.063, .045]). We repeated the analysis excluding overly regressive estimates, finding no support for degree of regression as associated with self-reported level of statistics training (*r*(1307) = -.009, *p* = .733, 95% CI [-.063, .045]).

## Comparing replication to target article’s findings

We summarized a comparison between the target article and the replication in Table 21.

###### Table 21 *Summary of statistical tests, effects, and evaluation of all six studies*

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Studies 1 and 2 (similarity and predictions)** | | | | | **Replication** | | | | | **Target article** |  | |
| **Analysis** | **DV** | | | | ***r*** | | ***p*** | **95% *CI*** | | **Effect** | **Interpretation** | |
| **a)** | Business administration description: likelihood — base rates | | | | -.081 | | .8 | [-.707, .616] | | *r* = .62 | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] | |
|  | Computer science description: likelihood — base rates | | | | .098 | | .8 | [-.605, .716] | | *r* = -.35 | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] | |
|  | Engineering description: likelihood — base rates | | | | .454 | | .2 | [-.301, .859] | | *r* = -.65 | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] | |
|  | Law description: likelihood — base rates | | | | .280 | | .5 | [-.472, .796] | | *r* = .33 | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] | |
|  | Library science description: likelihood — base rates | | | | -.436 | | .2 | [-.853, .321] | | *r* = -.03 | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] | |
|  | Medicine description: likelihood — base rates | | | | .102 | | .8 | [-.717, .603] | | *r* = .27 | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] | |
| **b)** | Business administration description: likelihood — similarity | | | | -.541 | | .2 | [-.878, .228] | | *r* = .88 | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] | |
|  | Computer science description: likelihood — similarity | | | | -.295 | | .4 | [-.802, .459] | | *r* = .96 | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] | |
|  | Engineering description: likelihood — similarity | | | | -.069 | | .9 | [-.701, .624] | | *r* = .97 | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] | |
|  | Law description: likelihood — similarity | | | | .797 | | .010 | [.282, .955] | | *r* = .93 | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] | |
|  | Library science description: likelihood — similarity | | | | .487 | | .2 | [-.261, .870] | | *r* = .88 | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] | |
|  | Medicine description: likelihood — similarity | | | | .729 | | .026 | [.126, .939] | | *r* = .92 | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] | |
| **c)** | Self-perceived accuracy — likelihood-base rate correlation coefficients | | | | .464\* | | .4 | [-.557, .927] | | null effect | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] | |
| **d)** | Null description: likelihood — base rates | | | | -.370 | | .3 | [-.830, .390] | | *r* = .74 | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] | |
|  |  | | | |  | |  |  | |  | [supported/not supported] | |
| *Note*. Pearson’s correlation, *df* = 7. \*: *df* = 4 | | | | | | | | | | | | |
| **Study 3 (prior vs individuating evidence)** | | | | | **Replication** | | | | | **Target article** | | |
|  | **Low engineer** | | **High engineer** | | ***t*** | ***df*** | ***p*** | ***d* and 95% *CI*** | | **Effect** | **Interpretation** | |
|  | *M* (%) | *SD* (%) | *M* (%) | *SD* (%) |  | |  |  | |  |  | |
|  | 49.91 | 1.708 | 51.65 | 3.795 | 0.865 | 4 | .435 | 0.387[-0.548, 1.279] | | *p* < .01, *d* = N/A\* | | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] |
|  |  |  |  |  |  |  |  |  | |  | [supported/not supported] | |
| *Note*. Paired samples t-test. \*: target article authors expected a null or very small effect but found support for an effect without reporting effect sizes. | | | | | | | | | | | | |
| **Study 4 (prediction vs evaluation)** | | | | | **Replication** | | | | | **Target article** | | |
| **Analysis** | **DV** | | | | ***F*** | ***df*** | | | ***p*** | **Effect** | | **Interpretation** |
| **a)** | Adjectives: SD in evaluation vs prediction | | | | 0.97 | (1, 653) | | | .32 | null effect | | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] |
|  | Reports: SD in evaluation vs prediction | | | | 0.05 | (1, 652) | | | .82 | null effect | | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] |
|  |  | | | |  |  | | |  |  | | [supported/not supported] |
| *Note*. Levene’s test of homogeneity | | | | | | | | | | | | |
| **Study 5 (prediction vs translation)** | | | **Replication** | | | | | | | **Target article** | | |
| **Analysis** | **Condition 1** | | | **Condition 2** | | | ***t*** | ***p*** | ***d* and 95% *CI*** | **Effect** | | **Interpretation** |
|  | Name | *M* | *SD* | Name | *M* | *SD* |  |  |  |  |  |  |
| **a) Mean predicted GPA** | Academic achievement | 1.965 | 0.448 | Mental concentration | 1.989 | 0.440 | -0.79 | .4 | -0.05[-0.19, 0.08] | null effect | | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] |
|  | Mental concentration | 1.989 | 0.440 | Sense of humor | 2.025 | 0.442 | 1.2 | .2 | -0.08[-0.22, 0.05] | Humor > mental\* | | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] |
| **b) SD predicted GPA** | Academic achievement | 1.394 | 0.216 | Mental concentration | 1.390 | 0.213 | 0.22 | .8 | 0.01[-0.12, 0.15] | null effect | | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] |
|  | Mental concentration | 1.390 | 0.213 | Sense of humor | 1.404 | 0.232 | -0.9 | .4 | -0.06[-0.19, 0.07] | Humor < mental | | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] |
| **c) r with predicted GPA** | Academic achievement | .002 | .333 | Mental concentration | .006 | .339 | -0.17 | .9 | -0.01[-0.14, 0.12] | null effect | | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] |
|  | Mental concentration | .006 | .339 | Sense of humor | -.000 | .345 | 0.26 | .8 | 0.02[-0.12, 0.15] | effect not supported | | N/A |
|  |  |  |  |  |  |  |  |  |  |  | | [supported/not supported] |
| *Note*. Independent samples t-test, *df* = 871 | | | | | | | | | | | | |
| **Study 6 (effect of consistency)** | | | **Replication** | | | | | | | **Target article** | | |
|  | **Correlated tests** | | **Uncorrelated tests** | | ***t*** | ***df*** | ***p*** | ***d* and 95% *CI*** | | **Effect** | **Interpretation** | |
|  | *M* | *SD* | *M* | *SD* |  | |  |  | |  |  | |
|  | 49.25 | 28.55 | 49.67 | 29.23 | -0.269 | 654 | .788 | -0.011[-0.087, 0.066] | | *p* < .001, *d* = N/A | | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] |
|  |  |  |  |  |  |  |  |  | |  | [supported/not supported] | |
| *Note*. Paired samples t-test | | | | | | | | | | | | |
| **Study 7 (persistence of non-regressive intuitions)** | | | | | **Replication** | |  | **Target article** | | |  |  |
| **Analysis** | **DV** | | | **S** | **R** | | **CR** | **S** | **R** | **CR** | **Interpretation** | |
| **a) Descriptive proportions** | Proportions 95% CI that were symmetric, regressive, or counter-regressive | | | 0% | 0% | | 100% | 67.59% 22.22% | | 10.19% | N/A (not a statistical test) | |
|  |  | **DV** |  | ***r*** | ***p*** | **95% *CI*** | |  |  |  |  |  |
| **b) Correlation** | Correlation between degree of regression and statistical knowledge | | | -.009 | .733 | [-.063, .045] | |  |  |  |  | [signal/no-signal] [consistent/inconsistent] [opposite/smaller/similar/larger] |
|  |  | | |  |  |  | |  |  |  |  | [supported/not supported] |
| *Note*. S = symmetric. R = regressive. CR = counter-regressive | | | | | | | | | | | | |

*Note*. The interpretation of outcomes is based on LeBel et al. (2019).

# General Discussion

[Planned for Stage 2: Following Dr./Prof. Regis Kakinohana comment regarding the need for mapping of the literature and replications, we will discuss the need for a follow-up systematic review of this prolific literature.]

[Planned for Stage 2: Following Dr./Prof. Peter Anthony White comment regarding the lacking definition and clear scope of representativeness, we will discuss the need to discuss and better define representativeness with clear measures and falsifiable criteria.]

[Planned for Stage 2: Following Dr./Prof. Peter Anthony White we will discuss the potential weaknesses and strengths of the unified design with our collecting all studies with the same sample in a with-in design, and our exploratory order analyses.]

[Planned for Stage 2: Following Dr./Prof. Peter Anthony White’s comment regarding the self-reported measure of confidence, we will discuss challenges, our findings, and future directions. In our peer review exchange we wrote the following as an initial base: “We conducted two large scale pre-registered replications, on MTurk (using CloudResearch) and Prolific, and both concluded very similar findings 45 years later (pre-registrations, materials, data, code, and reports available on https://doi.org/10.17605/OSF.IO/C3YVK). [...] To combat the possible explanation of this being a self-presentation bias, there are studies that show under-confidence in some studies with a similar methodology. The literature is nicely summarized in a recent book by Don Moore (2020) “Perfectly Confident: How to Calibrate Your Decisions Wisely”. ]

[Planned for Stage 2: Following Dr./Prof. Peter Anthony White’s comment: “I would suggest that, if they get results significant at .01 but not at .005, they could discuss these or at least list them, so that readers could get a feel for whether there is any likelihood of type 2 errors”. We will discuss replications of problems in which the findings fell in between .05 and our set alpha (.005/.001)]

# Conclusion

[To be completed in Stage 2 following data collection]

# References

Au, N., & Feldman, G. (2020). Revisiting “Goals as Reference Points”: Replication and extensions of Heath et al. (1999). <https://doi.org/10.17605/OSF.IO/WMQTB>

Baratgin, J., & Noveck, I. A. (2000). Not only base rates are neglected in the Engineer–Lawyer problem: An investigation of reasoners' underutilization of complementarity. *Memory & Cognition, 28*(1), 79–91. [https://doi.org/10.3758/BF03211578](https://psycnet.apa.org/doi/10.3758/BF03211578)

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. doi:10.3758/bf03193146

Feldman, G. (2023). Registered Report Stage 1 manuscript template. <https://doi.org/10.17605/OSF.IO/YQXTP>

Fischhoff, B., & Bar-Hillel, M. (1984). Diagnosticity and the base-rate effect. *Memory & Cognition, 12*(4), 402–410. [https://doi.org/10.3758/BF03198301](https://psycnet.apa.org/doi/10.3758/BF03198301)

Gigerenzer, G., Hell, W., & Blank, H. (1988). Presentation and content: The use of base rates as a continuous variable. *Journal of Experimental Psychology: Human Perception and Performance, 14*(3), 513–525. [https://doi.org/10.1037/0096-1523.14.3.51](https://psycnet.apa.org/doi/10.1037/0096-1523.14.3.513)3

Hong, C. & Feldman, G. (2023). Revisiting the “Belief in the law of small numbers”: Conceptual replication and extensions Registered Report of problems reviewed in Tversky and Kahneman (1971). Stage 1 In-principle acceptance from Peer Community in Registered Report. <https://doi.org/10.17605/OSF.IO/MNS7J>

Jané, M., Xiao, Q., Yeung, S., Ben-Shachar, M. S., Caldwell, A., Cousineau, D., Dunleavy, D. J., Elsherif, M., Johnson, B., Moreau, D., Riesthuis, P., Röseler, L., Steele, J., Vieira, F., Zloteanu, M., & Feldman, G. (2024). Guide to Effect Sizes and Confidence Intervals. <http://dx.doi.org/10.17605/OSF.IO/D8C4G>

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 49–81). Cambridge University Press. <https://doi.org/10.1017/CBO9780511808098.004>

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251. [https://doi.org/10.1037/h003474](https://psycnet.apa.org/doi/10.1037/h0034747)7.

Koehler, J. J. (1996). The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. *Behavioral and Brain Sciences, 19*(1), 1–53. [https://doi.org/10.1017/S0140525X00041157](https://psycnet.apa.org/doi/10.1017/S0140525X00041157)

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*, 389-402.

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A brief guide to evaluate replications. *Meta-Psychology*, 3, 1-9. <https://doi.org/10.15626/MP.2018.843>

Li, M. Y. , & Feldman, G. (2022). Revisiting mental accounting classic paradigms: Replication of the problems reviewed in Thaler (1999). Stage 1 In-principle acceptance from Peer Community in Registered Report. Stage 2 preprint retrieved from: <https://doi.org/10.17605/OSF.IO/V7FBJ>

Mayiwar, L., Wan, K. H., Løhre, E., & Feldman, G. (2024). Revisiting representativeness classic paradigms: Replication and extensions of the problems reviewed in Kahneman and Tversky (1972). <https://doi.org/10.17605/OSF.IO/NHQC4>

Morsanyi, K., & Handley, S. (2012). Does thinking make you biased? The case of the engineers and lawyers problem. *Proceedings of the annual meeting of the cognitive science society, 34*(34). <https://escholarship.org/uc/item/6gt9d86j>

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73(1), 719-748.

Novemsky, N., & Kronzon, S. (1999). How are base-rates used, when they are used: A comparison of additive and Bayesian models of base-rate use. J*ournal of Behavioral Decision Making, 12*(1), 55–67. [https://doi.org/10.1002/(SICI)1099-0771(199903)12:1<55::AID-BDM306>3.0.CO;2-M](https://psycnet.apa.org/doi/10.1002/(SICI)1099-0771(199903)12:1%3C55::AID-BDM306%3E3.0.CO;2-M)

Olsson-Collentine, A., van Assen, M. A. L. M., & Wicherts, J. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*. <https://doi.org/10.1037/bul0000294>

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22-27. <https://doi.org/10.1016/j.jbef.2017.12.004>

Patil, I. (2021). Visualizations with statistical details: The 'ggstatsplot' approach. *Journal of Open Source Software*, *6*(61), 3167.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*(5), 407-413. <https://doi.org/10.1111/j.1467-9280.2006.01720.x>

Petrov, N. B., Chan, Y. K. M., Lau, C. N., Kwok, T. H., Chow, L. C. E., Lo, W. Y., Song, W., & Feldman, G. (2023). Sunk Cost Effects for Time Versus Money: Replication and Extensions Registered Report of Soman (2001). *International Review of Social Psychology,* 36(1): 17, 1–18. DOI: https://doi. org/10.5334/irsp.883

R Core Team (2022). R: A Language and environment for statistical computing. (Version 4.1) [Computer software]. Retrieved from <https://cran.r-project.org>. (R packages retrieved from CRAN snapshot 2023-04-07).

The JAMOVI project (2023). JAMOVI. (Version 2.4) [Computer Software]. Retrieved from <https://www.jamovi.org>.

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. *Cognitive Psychology, 5*(2), 207–232. [https://doi.org/10.1016/0010-0285(73)90033-9](https://psycnet.apa.org/doi/10.1016/0010-0285(73)90033-9)

Vonasch, A. J., Hung, W. Y., Leung, W. Y., Nguyen, A. T. B., Chan, S., Cheng, B. L., & Feldman, G. (2023). ‎“Less Is Better” in Separate Evaluations Versus “More Is Better” in Joint Evaluations: Mostly ‎Successful Close Replication and Extension of Hsee (1998)‎. *Collabra: Psychology*, 9(1). doi: 10.1525/collabra.77859 <https://doi.org/10.1525/collabra.77859>

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. [https://ggplot2.tidyverse.org](https://ggplot2.tidyverse.org/).

Yeung, S. K., & Feldman, G. (2022). Revisiting the temporal pattern of regret in action versus inaction: Replication of Gilovich and Medvec (1994) with extensions examining responsibility. *Collabra: Psychology*, *8*(1), 37122. <https://doi.org/10.1525/collabra.37122>

Zhu, M., & Feldman, G. (2023). Revisiting the Links Between Numeracy and Decision Making: Replication Registered Report of Peters et al.(2006) With an Extension Examining Confidence. *Collabra: Psychology*, *9*(1). <https://doi.org/10.1525/collabra.77608>

1. We searched Google Scholar for replications of Kahneman and Tversky (1973) using the “search within citing articles” function. The following keywords were used: “replicat\*”, “review”, “freshmen”, “personality sketch”, “lawyers”, “engineers”, “adjectives”, “evaluation”, “translation”, “consistency”, and “aptitude test”. [↑](#footnote-ref-2)
2. We note that in some instances, Prolific recruits participants beyond the specified sample size. This is due to the platform sometimes incorrectly classifying valid completed responses as ‘timed out’ or ‘returned’. We will not exclude any complete valid responses from our dataset, and will include any additional completed responses obtained from Prolific. We will only conduct our data analysis once all data has been collected. [↑](#footnote-ref-3)
3. Original authors reported that participants were recruited by means of a student paper from the University of Oregon (see footnote 3 in p. 238), but did not further specify sample characteristics. [↑](#footnote-ref-4)