Misprediction of affective outcomes due to different evaluation modes: Replication and extension of two distinction bias experiments by Hsee and Zhang (2004)

*Farid Anvari[a], *Jerome Olsen[b], *Wing Yiu Hung[c], and ^*Gilad Feldman[c]

[a] Strategic Organization Design group, Department of Marketing and Management, University of Southern Denmark, Campusvej 55, 5230 Odense, Denmark.
faridanvari.phd@gmail.com

[b] Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53113, Bonn, Germany.
olsen@coll.mpg.de

[c] University of Hong Kong, Department of Psychology, The University of Hong Kong, Pokfulam Road, Hong Kong.
gfeldman@hku.hk

In press at Journal of Experimental Social Psychology

Accepted for publication on September 6, 2020

*Contributed equally, joint first author
^Corresponding author
Word: abstract – 249, manuscript - 7358

Authorship Declaration

Hung Wing Yiu conducted the replication as part of her undergraduate thesis in psychology (PSYC4008), as part of a Bachelor of Social Sciences Degree at the University of Hong Kong.

The corresponding author, Gilad Feldman, was the thesis advisor for Hung Wing Yiu. Gilad supervised each step in the project, conducted the pre-registrations, ran data collection, and provided guidance on revisions of the manuscript.

Farid Anvari coordinated and drafted the manuscript, conducted draft revisions and editing, and consulted with Jerome Olsen on statistical analyses and data visualizations.

Jerome Olsen conducted the statistical analyses and data visualizations, and consulted with Farid Anvari on coordinating and drafting the manuscript, and conducted draft revisions and editing.

Farid, Jerome, and Gilad finalized the manuscript for submission.

The current replication is part of the larger ‘mass pre-registered replications in judgment and decision-making’ project lead by Gilad Feldman. The project aims to revisit well known research findings in the area of judgment and decision making (JDM) and investigate the replicability of these findings. As part of the initiative the students engage in pre-registered replications and extensions to revisit well-known findings as part of regular one-semester coursework and/or thesis.

Declaration of Conflict of Interest

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

Financial Disclosure/Funding:

None.

Contributor Roles Taxonomy

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

Role	Hung	Farid	Jerome	Gilad
Conceptualization	X			X
Pre-registration	X			X
Data curation	X		X	X
Formal analysis			X	X
Funding acquisition				X
Investigation	X			X
Pre-registration peer review / verification				X
Data analysis peer review / verification				X
Methodology	X			X
Project administration				X
Resources				X
Software				X
Supervision				X
Validation		X	X	X
Visualization			X	X
Writing-original draft		X		
Writing-review and editing		X	X	X

Abstract

Hsee and Zhang (2004) argued that when people face a decision and must predict future affective states, they are often in a joint evaluation (JE) mode where direct comparisons between different choice options are relatively easy. When actually experiencing (or predicting affective states for) only one of these options, people are usually in single evaluation (SE) mode where direct comparisons are more difficult. This situational difference in evaluation mode was observed to lead to overpredictions of positive affect (happiness) when people were in JE in comparison to SE for options that were quantitatively different, but there were no overpredictions for options that were qualitatively different. This effect was coined distinction bias. In the present paper, we replicated (and extended on) Studies 1 and 2 from Hsee and Zhang with 824 MTurk participants. In Study 1 we replicated the original findings: Relative to people in SE, people in JE overpredicted the happiness derived from quantitatively different hypothetical scenarios (i.e., selling 80 vs 160 books, or 160 vs 240 books), but did not overpredict the difference between qualitatively different hypothetical scenarios (i.e., selling 0 books vs 80 books; 0 being the implicit reference point that makes the two scenarios qualitatively different). Study 2 failed to find support for the original findings: People in JE did not consistently overpredict happiness for quantitatively different scenarios (i.e., copy-pasting 25 negative words vs 10 negative words, or 10 positive words vs 25 positive words). Taken together, the present paper provides mixed support for distinction bias.

Keywords: distinction bias, evaluation mode, affective forecasting, decision making, experienced utility, misprediction

Misprediction of affective outcomes due to different evaluation modes: Replication and extension of two distinction bias experiments by Hsee and Zhang (2004)

Introduction

We often decide and choose between the options available to us based on what we predict would produce the greatest utility or happiness as one manifestation of utility. For example, in deciding whether to get one scoop of ice cream or two (or even three), whether to take the job that pays more but is further from home, whether to pay extra for a larger apartment or T.V., we make predictions for which alternative would make us happiest (i.e., we engage in affective forecasting). But are such predictions accurate? Rational choice theories assume that people know and choose the option which would give them the greatest happiness (Scott, 2000). However, researchers have argued that people's predictions for what will make them happy (predicted utility), and what they choose (decision utility), can systematically differ from what actually makes them happy (experienced utility; e.g., Kahneman, 2000; Markman & Hirt, 2002; Wilson & Gilbert, 2003).

Research suggests that people indeed make mispredictions with respect to what would (vs what does) make them happiest, and various reasons for this bias have been proposed and identified (e.g., Dunn et al., 2003; Gilbert et al., 2002; Kahneman & Snell, 1992; Kurtz et al., 2007; Morewedge et al., 2010; Novemsky & Ratner, 2003; O'Brien & Roney, 2017; Ratner et al., 1999; Schkade & Kahneman, 1998; for a comprehensive outline of when, how, and why people make mispredictions in affective forecasting, see Wilson & Gilbert, 2003). One such explanation for misprediction in affective forecasting was proposed by Hsee and Zhang (2004). They argued that predictions and choices are often made in an evaluation mode that typically differs from the evaluation mode people are in during experience. When predicting or choosing, people tend to be in joint evaluation (JE) mode such that they make direct comparisons between multiple options. When

experiencing, or predicting an experience about one event in isolation, people tend to be in single or separate evaluation (SE) mode. Hsee and Zhang proposed (and in 3 studies found support for the idea) that people in JE may, in certain situations, overpredict the differences in affective consequences between options compared to people in SE. This was called distinction bias.

In the present paper, we report replications and extensions of Studies 1 and 2 from Hsee and Zhang (2004). The remainder of the paper is as follows. First, we describe the theoretical underpinnings of distinction bias and provide details of Studies 1 and 2 from Hsee and Zhang which provided support for the phenomenon. Second, we outline the need for replicating the original studies. Third, we detail how our replication differs from the original studies, including how we extend on the original findings. And, finally, we report and discuss the findings of our replication studies and how they compare to the original findings.

Theory of Distinction Bias and the Original Studies

Hsee and Zhang (2004) proposed that people will tend to be in JE mode when they are predicting how happy some future experience will make them. In JE mode, people are typically presented with multiple options that have different values of an attribute so that they can easily compare and differentiate the desirability of the different values—that is, they can easily compare the alternative options based on the amount of the attribute they have relative to one another. Therefore, when predicting the affective outcomes of different options, people in JE mode predict greater (lesser) happiness for subsequently increasing values of a desired (undesired) attribute.

On the other hand, Hsee and Zhang (2004) proposed that when experiencing something, people are often likely to be in SE mode. In SE mode, people typically experience only one value of the attribute—that is, they are presented with only one option—so that they cannot easily compare it with alternative options that have different values of the attribute. Therefore, people in SE mode do not have a precise idea of how (quantitatively) positive or negative the isolated value is, though they

may be able to generally evaluate it as (qualitatively) positive or negative in relation to a reference point. As Hsee and Zhang explained, values of an attribute are considered to be quantitatively different from each other if they fall on the same side of a reference point, whereas qualitatively different values cross the reference point (i.e., either one value falls on the reference point or the two values fall on either side of it). For example, whether someone sells 80 books or 240 is a quantitative difference, whereas selling no books is qualitatively different to selling 80 books, assuming a reference point of 0 books sold. However, the reference point can vary. For example, if someone has expectations that they will sell 160 books then this becomes a new reference point, and the difference between selling 80 or 240 books is a qualitative difference since they now fall on opposing sides of the new reference point. As such, when people are in SE mode, the happiness predicted or experienced for different options does not increase (decrease) linearly with increasing values of a desired (undesired) attribute but, instead, depends on whether the options are qualitatively evaluated as either positive or negative.

Distinction bias occurs when people overpredict how much happier they would be when comparing one value of an attribute with another. Specifically, people in JE mode tend to overestimate the difference in happiness elicited by quantitatively (but not qualitatively) different options in comparison to people in SE. In other words, when people in JE are asked to predict the happiness between two people presented with two different options, they overpredict the difference when the different options are quantitatively different but not when they are qualitatively different. This is because the evaluation function of an attribute systematically differs depending on whether the evaluation is made in JE or in SE. People in JE observe the alternative values of an attribute and can thus compare and differentiate the different options quite easily, in terms of their relative desirability (Kleinmuntz & Schkade, 1993; Tversky, 1969). As a result, the evaluation function in JE is smooth and steep (see solid line in Figure 1). In contrast, people in SE only observe a single

value of the attribute and, therefore, cannot easily compare alternative values in terms of desirability. The only exception is that for qualitatively different values of the attribute people in SE can vaguely note that a value is good if it is positive or above a reference point and bad if it is negative or below a reference point. The evaluation function in SE (see also, Hsee et al., 1999) will thus be steep around the reference point and flat elsewhere (see dashed line in Figure 1). Therefore, people in SE will predict and experience greater happiness from an option that is qualitatively more positive than another, but for quantitatively different options people will predict and experience similar levels of happiness. In sum, Hsee and Zhang (2004) proposed that distinction bias is the result of (1) the different evaluation modes adopted when people are predicting and experiencing, and (2) the difference in the nature of the options (quantitative vs qualitative differences).

Hsee and Zhang (2004) presented three studies supporting the theory of distinction bias. In the present paper we focus on Studies 1 and 2, which investigated distinction bias in the context of mispredictions. Study 3 focused on mischoice in the context of tradeoffs and is not the focus of the present paper.

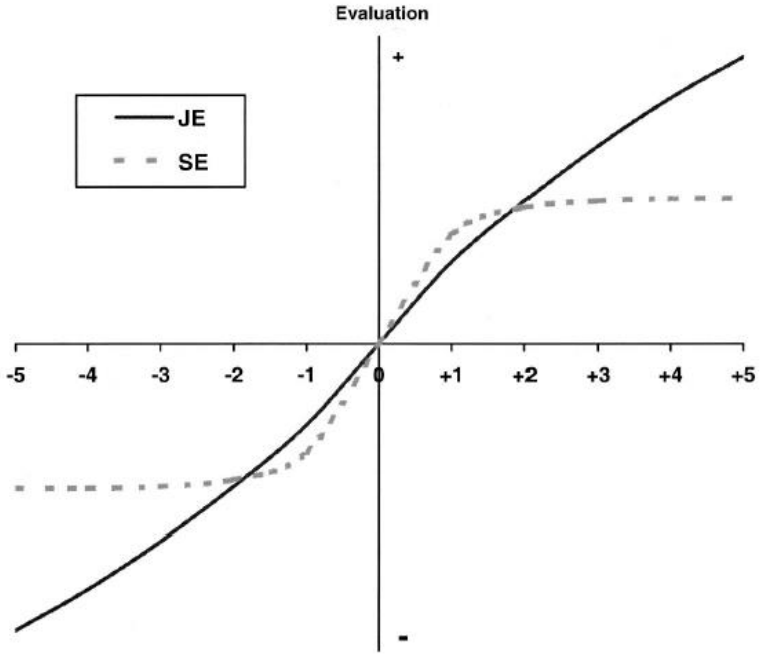

Figure 1. The SE curve (for a hypothetical attribute) is flatter than the JE curve except around a reference point (zero in this case). Reprinted from "Distinction bias: Misprediction and mischoice due to joint evaluation", C.K. Hsee, & J. Zhang, 2004, *Journal of Personality and Social Psychology, 86*(5), 680-695. Copyright 2004 by the American Psychological Association. Reprinted with permission.

Original Study 1. In Study 1, Hsee and Zhang (2004) allocated participants to be in one of five conditions (one condition for JE mode and four conditions of SE mode). All participants were told to imagine that their favorite hobby was writing poems and that they were going to try selling copies of a book of poems they had written. Participants in JE mode were asked to consider four scenarios (no one, 80, 160, or 240 people bought the book), imagine that it had happened to them, and to predict how happy (from *1 = extremely unhappy*, to *9 = extremely happy*) they would be in each of those situations individually. Participants in the four SE conditions (corresponding to the four scenarios) were given only one of the scenarios, told to assume the scenario had happened, and then asked to rate how happy they would be on the same scale. Hsee and Zhang argued that, whereas the differences between 80, 160, and 240 people buying the book are quantitative, there is a qualitative difference between no one buying the book and 80 people buying the book (assuming a natural reference point of 0 buyers). Hsee and Zhang hypothesized that participants in JE would

overpredict (compared to those in SE) the difference in happiness between the 80, 160, and 240 buyers scenarios, but that they would be less likely to make such an overprediction between the no-buyers and the 80-buyers scenarios.

The results supported the hypotheses. Participants in JE predicted significantly greater happiness for each subsequent increase in the number of buyers for the quantitatively different scenarios (i.e., 80 vs 160 buyers, and 160 vs 240 buyers), whereas participants in SE did not give significantly different happiness ratings for these scenarios. In contrast, participants in JE predicted significantly greater happiness in the 80 buyers scenario than the no buyers scenario (the qualitatively different scenarios), which was in line with the SE ratings which also showed significantly greater happiness in the 80 buyers scenario than the no buyers scenario. In sum, although participants in JE overpredicted the difference in happiness for the quantitatively different values, they did not make an overprediction for the qualitatively different values. The descriptive statistics of these findings are presented alongside the descriptive results of the replication findings further below, for ease of comparison (see Table 1); for inferential statistics see Table S1.

Original Study 2. In Study 2, Hsee and Zhang (2004) conceptually replicated the findings from Study 1 but included additional SE conditions in which participants actually experienced one option and reported their happiness. Thus, Study 2 consisted of comparing predictions in JE with real experience in SE. The study also included a comparison with predictions in SE. The focal interest seemed to be on the comparison of JE predicted-experience and SE real-experience. Hsee and Zhang included this comparison to rule out that distinction bias only applies to predictions and does not translate to experiences in SE.

The JE group was asked to imagine that an experimenter had asked them to do one of four tasks: read a list of 25 negative words, such as *hatred* and *loss*; read a list of 10 negative words, such as *hatred* and *loss*; read a list of 10 positive words, such as *love* and *win*; read a list of 25

positive words, such as *love* and *win*. The JE group was asked to compare and predict how they would feel if they did each of these tasks. The four SE real-experience groups were given only one of the four tasks to actually perform and then to rate how they felt. The four SE predicted-experience groups, instead of actually performing it, were asked to predict how they would feel if they did the one task presented to them. All groups gave ratings of happiness on the same scale as in Study 1. Hsee and Zhang (2004) proposed that the differences between 25 and 10 negative words and between 10 and 25 positive words were quantitative, whereas the difference between 10 positive words and 10 negative words was qualitative due to the change in valence. Thus, they hypothesized that people in JE would overpredict the differences in happiness between 25 and 10 negative words and between 10 and 25 positive words, but not for the difference between 10 negative words and 10 positive words (compared to both SE real-experience and SE predicted-experience).

The hypotheses were supported. The JE group predicted significantly more happiness (or less unhappiness) for 10 negative words compared to 25 negative words and significantly more happiness for 25 positive words compared to 10 positive words. However, these differences were not significant for the SE real-experience and SE predicted-experience groups.[1] On the other hand, for reading 10 positive words compared to 10 negative words, the JE group predicted significantly more happiness and this difference was also significant for the SE real-experience and SE predicted-experience groups. In sum, people in JE overpredicted the difference in happiness when the tasks differed quantitatively (i.e., in degree) but not when the tasks differed qualitatively (i.e., in

[1] The original study by Hsee and Zhang (2004) did not report inferential statistics for the SE predicted-experience group, though they presented the results in a line graph without error bars. Visually, results of the SE predicted-experience and SE real-experience groups seemed rather similar.

valence). The means and standard deviations for the original Study 2 are presented alongside the replication's descriptive results (see Table 2 further below); for inferential statistics see Table S1.

Need for Replication

We chose to replicate the Studies (1 and 2) by Hsee and Zhang (2004) based on two factors: the absence of direct replications and impact (Coles et al., 2018; Isager, 2019). Multiple large-scale replication projects in recent years have revealed that replication rates are often far from ideal (e.g., Camerer et al., 2016; Hagger et al., 2016; Klein et al., 2014, 2018; Open Science Collaboration, 2015). Replication contributes to the credibility of research findings and the building of a cumulative knowledge base (Brandt et al., 2014; Etienne P. LeBel et al., 2018; Rosenthal, 1990). It is thus an important aspect of all scientific fields. Direct, or close, replications (i.e., using the same operationalizations of the dependent and independent variables) are important because only replications that are sufficiently similar to the original study can help us update the strength of support for the original hypothesis—without such similarity, studies can only speak to the generalizability of the phenomenon, not its replicability (LeBel et al., 2018; LeBel et al., 2019). Although there have been conceptual replications of the original studies (e.g., Dunn et al., 2017; Hsee et al., 2013), we are unaware of any close replications that can speak to the original hypothesis. Given that publication bias (the phenomenon that statistically significant results in support of hypotheses are more likely to be published than nonsignificant results and thus) could hide nonsignificant findings from the published record (Bakker et al., 2012), a close replication of the original studies adds value by providing evidence that can either strengthen or weaken support for the original distinction bias hypothesis.

Moreover, Hsee and Zhang's (2004) paper is published in one of the most visible personality and social psychology journals and has attracted significant scientific interest with over 320 citations (as of May 2020, Google Scholar). This research interest has extended to applied research

fields such as consumer behavior (e.g., Shen et al., 2012) and personnel selection (e.g., Brooks et al., 2009). Importantly, Hsee and Zhang's (2004) findings form the foundation for several components of an important and widely encompassing theory of value sensitivity, the General Evaluability Theory (Hsee & Zhang, 2010), which has itself been highly cited (304 citations as of May 2020, Google Scholar). As such, a close replication that tests the reliability of support for the original hypothesis will be a valuable contribution to the field.

Overview of the Present Studies

We conducted close replications of Studies 1 and 2 from Hsee and Zhang (2004). We note that there are some differences between the present replication studies and the original studies. The differences that we introduced, which are detailed below, are either incidental to conducting the replications online or designed to extend on the original findings. Using the classification system proposed by LeBel et al. (2018), our studies may be considered as very close replications of the original studies (see Table S2, Supplemental Materials). (See Table S3 for a comparison, between the original and replication studies, of the wording for the hypotheses.)

Adjustments to the original study

First, whereas participants in the original studies were university students from the U.S., in the replication studies we used Amazon Mechanical Turk (MTurk) participants who were U.S. residents aged 18 years or over. Second, whereas the original studies were presumably conducted in the lab, the replication studies were conducted online via the Qualtrics survey software. Research shows high levels of consistency between original lab-based samples and more recent online samples, with little variability in results being attributable to whether studies are conducted in lab versus online (e.g., Klein et al., 2018; Ziano et al., 2019). Third, the original Study 2 involved participants reading words, but, because the replication studies were conducted online, the

replication involved also copying and pasting the words (we made this adjustment in an attempt to ensure the words were being attended to). We consider these to be incidental differences because they are not critical to the theory of distinction bias. For example, distinction bias is not a theory about university students but about people more generally and so it should apply to people on MTurk. For comparability, we present sample characteristics of the original and replication studies in Table S4 (see Supplemental Materials).

Extensions

Apart from the incidental differences described above, we wanted to extend on the original findings. The main way in which we extended on the original findings was by introducing an alternative reference point for what would be considered a qualitative (as opposed to a quantitative) difference. To do this, in Study 1, we included an additional JE group and two additional SE groups. Specifically, participants in these three extension conditions were told to imagine that they had expected that 160 people would buy their poem book. Hence, the reference point of 160 would make the difference between 80 and 240 buyers a qualitative difference (given that the values would now fall on either side of the reference point). Based on the theory, we therefore expected that people in the JE extension would not overpredict the difference between 80 and 240 buyers scenarios. In other words, participants in the JE extension would predict greater happiness for 240 buyers as compared to 80 buyers, as would participants in the SE extension.

Hence, whereas in the original design the assumed reference point was 0 buyers, such that people in JE were expected to overpredict the difference between 80 and 240 buyers in comparison to people in SE, in the extension we included two conditions that changed the reference point to 160 buyers such that people in JE would not overpredict the difference between 80 and 240 buyers.

Another exploratory extension was to examine whether people in JE differed in their level of certainty when making predictions for quantitatively different values of an attribute as compared to

making predictions for qualitatively different values. The results of this extension were inconclusive. To maintain the clarity and conciseness of the present paper, we introduce and report the results about certainty in the Supplemental Materials.

Preregistration, Open Data/Materials, and Open Reporting Statement

For both studies, the sample size, conditions, and measures that we report were all preregistered, as were the analyses. See preregistration here: https://osf.io/8hr4t?view_only=9f636f168e434e9e8ad453c6b72b32c9 . Specifically, to compare JE predictions we preregistered the use of paired samples *t*-tests, and to compare SE groups the use of independent samples *t*-tests. We report all deviations from the preregistration due to oversights in the plan (i.e., power analysis and proposed use of equivalence testing) in Table S5 (see Supplemental Materials). All data and materials for the two studies that we report can be found here https://osf.io/x6cq9/?view_only=42abb1c032f444d19f2dba0d834b5106 .

For both studies, we report all measures, manipulations, and exclusions, as well as the method of determining the final sample size. Data collection was complete before analyses were conducted. We report effect sizes, exact p-values, means, standard deviations, and inter-variable correlations when relevant to the analyses. We use Bonferroni corrections (decided upon post-hoc) to control error rates for multiple comparisons. Studies 1 and 2 involve 8 and 9 *t*-tests, respectively. Therefore, the corrected alpha levels for Studies 1 and 2 are .0063 and .0055, respectively.

Participants

The participants took part in both Studies 1 and 2 which were presented in random counterbalanced order.[2] We planned to recruit at least 773 participants via Amazon Mechanical

[2] There was no significant effect of presentation order on the results. See pp. 14-19 of the Supplemental Materials for details.

Turk (MTurk). This initial targeted sample size was determined through a suboptimal power analysis (see Table S5, Supplemental Materials). Nevertheless, we report improved power determination analyses with power curves that illustrate statistical power of our sample for various effect sizes (see Figure 2). We note that the resulting sample size was twice the original sample size in each cell (see Table 1 and Table 2).

A total of 824 participants (373 male, 451 female, Age: *M* = 40.2 years, *SD* = 12.0 years) were recruited, resulting in at least 100 participants per condition in Study 1 and 86 per condition in Study 2. We preregistered our intention to use the full sample in the analyses and to conduct supplementary analyses based on exclusion criteria outlined in the preregistration plan. Exclusions had little effect on the results (see Table S6, Supplemental Materials) and qualitative interpretations were unaffected. We therefore report our analyses on the full sample for both studies.

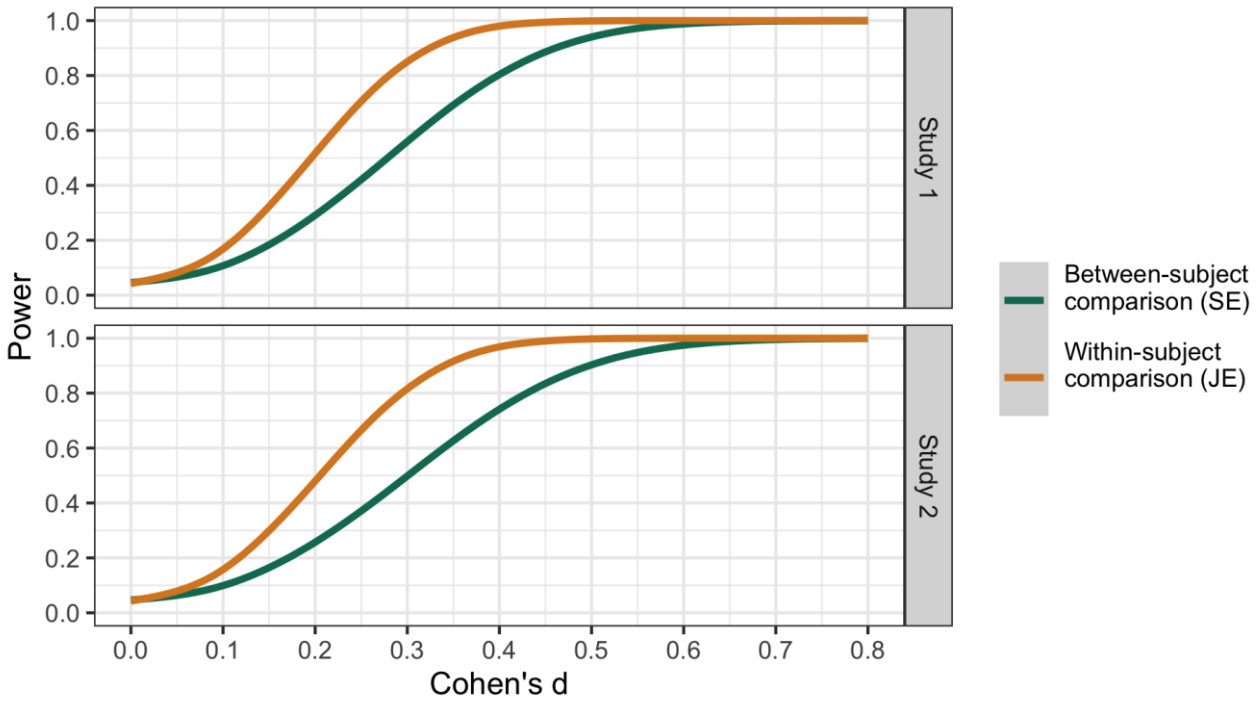

Figure 2. Power curves from power determination analyses. The within-subject comparison lines represent the power curves for the JE comparisons and the between-subject comparison lines represent the power curves for SE comparisons.

Study 1 (Poem Book)

Methods

Materials, measures, and procedure. All participants were asked to imagine that their favorite hobby was writing poems and that they were trying to sell a book of their poems on campus. Participants were randomly assigned to one of eight groups (one JE group, four SE groups, one JE extension group, and two SE extension groups).

JE group. Participants in the original joint evaluation (JE) group were presented with all of the four scenarios:

So far no one has bought your book

So far 80 people have bought your book

So far 160 people have bought your book

So far 240 people have bought your book

They were then presented with each scenario in randomized order and asked to rate how they would feel if the scenario had happened. Ratings were given on a 9-point scale (from *1 = extremely unhappy*, to *9 = extremely happy*).

SE group. Participants in the four original separate evaluation (SE) groups were presented with only one of the options from above, and asked to rate how they would feel on the same 9-point happiness scale. For example, in one group, participants were asked to imagine that no one had bought the book (without being presented with any of the other scenarios).

JE extension group. In the extension JE group, participants were presented with the above four scenarios with an additional statement, "it was predicted that 160 people would buy your book". Next, they were presented with two of the four scenarios, in random order, and asked to rate how they would feel on the 9-point happiness scale. Specifically, they were asked how they would feel if 80 people bought the book and how they would feel if 240 people bought the book.

SE extension group. Participants in the remaining two extension SE groups were presented with either the scenario in which there were 80 buyers or 240 buyers, with an additional statement, "it was predicted that 160 people would buy your book". They were then asked to rate how they would feel on the 9-point happiness scale.

Results and Discussion

The descriptive statistics for the ratings on predicted happiness for each scenario partitioned by condition, for Study 1, are presented in Figure 3 and Table 1. To analyze the data, we used Student's paired samples *t*-tests for comparisons between scenarios for the JE group and Welch's independent samples *t*-tests for the comparisons between the SE groups. All tests were two-tailed. According to the theory of distinction bias, compared to people in SE, people in JE were expected to (1) overpredict the difference in happiness between the 80, 160, and 240 buyers scenarios (i.e., the quantitatively different scenarios), (2) but not expected to overpredict the difference between the 0 and 80 buyers scenarios (i.e., the qualitatively different scenarios). The pattern of results supports the distinction bias hypotheses.

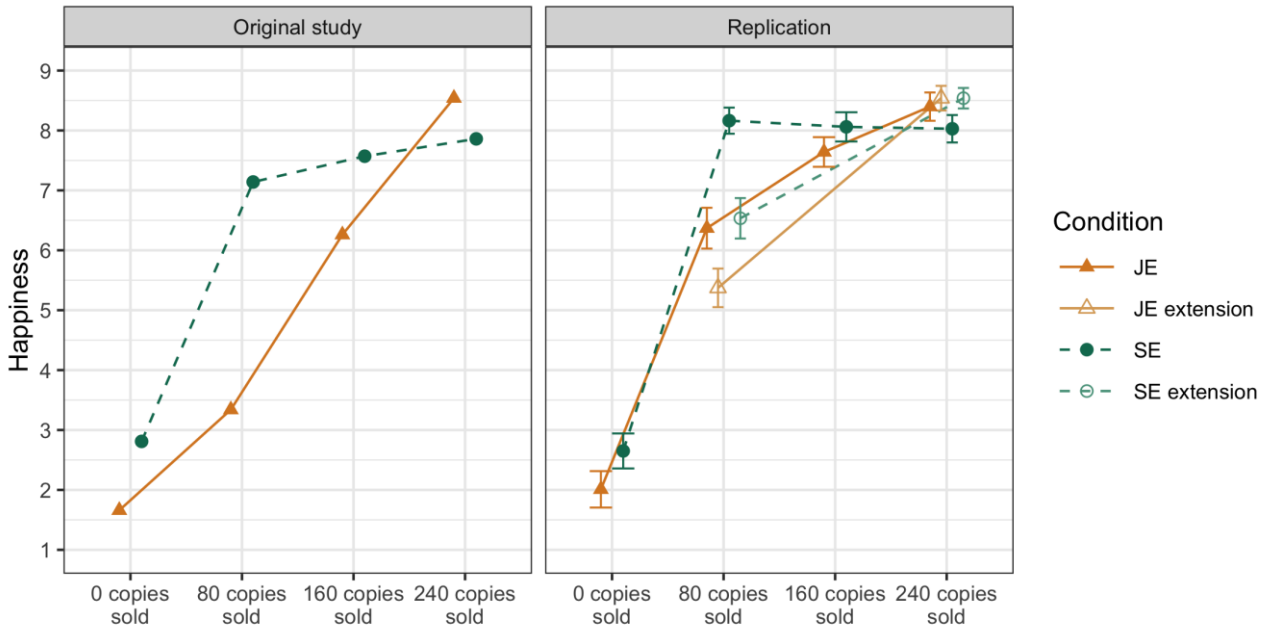

Figure 3. Poem book study: People in joint evaluation (JE) overpredicted the differences in happiness ratings for quantitatively difference scenarios (80 vs 160 vs 240 buyers), compared to people in separate evaluation (SE), but underpredicted the difference for the qualitatively different scenarios (0 vs 80 buyers). JE extension and SE extension refer to the evaluation modes when the reference point was explicitly stated to be 160 buyers, so that the 80 and 240 buyers scenarios in the extension are now qualitatively different.

Table 1
Descriptive Statistics for Study 1 from the Original and Replication Studies

Evaluation mode	Number of books sold	Original study		Replication study		
		N	*M*	*N*	*M*	*SD*
JE	0	249/5	1.66	103	2.01	1.57
JE	80		3.34		6.37	1.77
JE	160		6.26		7.64	1.28
JE	240		8.54		8.40	1.22
SE	0	249/5	2.81	103	2.65	1.51
SE	80	249/5	7.14	104	8.16	1.13
SE	160	249/5	7.57	100	8.06	1.25
SE	240	249/5	7.86	105	8.03	1.20
JE ext.	80			103	6.53	1.75
JE ext.	240				8.54	0.89
SE ext.	80			102	5.37	1.66
SE ext.	240			102	8.54	1.06

Note. For the original study only the overall sample size was reported. SDs were not reported at all.

Distinction bias replication results. Comparing the quantitatively different scenarios, people in JE predicted statistically significant lower happiness for the 80 buyers scenario than the 160 buyers scenario ($t_{(102)} = 11.08$, $p < .001$, $d = 0.77$, $CI_{95\%}[0.61, 0.92]$)[3], and for the 160 buyers scenario than the 240 ($t_{(102)} = 6.01$, $p < .001$, $d = 0.6$, $CI_{95\%}$ [0.39, 0.82]). In contrast, when comparing people in SE in the quantitatively different scenarios, the difference between the 80 and 160 buyers scenarios was not statistically significant ($t_{(198.45)} = -0.62$, $p = 0.536$, $d = 0.09$, $CI_{95\%}$[-0.19, 0.36]), nor was the difference between the 160 and 240 buyers scenarios ($t_{(201.4)} = -0.18$, $p =$

[3] The effect sizes we report for JE comparisons are Cohen's d_{av}, because it is an effect size measure for within-subjects designs which is generalizable (and thus comparable) to effect sizes in between-subjects designs (Lakens, 2013).

.854, d = 0.03, $CI_{95\%}$[-0.25, 0.3]). Importantly, to establish that people in JE actually overpredicted the increases in happiness for the quantitatively different scenarios, compared to people in SE, we need to examine whether the confidence intervals of the effect sizes for the differences between the two conditions in JE and SE overlap. As can be seen, the confidence intervals for people in JE when comparing 80 with 160 buyers (i.e., [0.61, 0.92]) and when comparing 160 and 240 (i.e., [0.39, 0.82]) buyers did not overlap with the confidence intervals for the corresponding SE comparisons (i.e., [-0.19, 0.36] and [-0.25, 0.3], respectively). Therefore, people in JE overpredicted the difference between the 80, 160, and 240 buyers scenarios.

In comparing the 0 and 80 buyers scenarios (i.e., the qualitatively different values), we found that people in JE predicted significantly higher happiness for the latter ($t_{(102)}$ = 21.66, p < .001, d = 2.6, $CI_{95\%}$[2.11, 3.1]) and comparing the 0 buyers SE group with the 80 buyers SE group also showed that the latter predicted significantly higher happiness ($t_{(189.04)}$ = 29.65, p < .001, d = 4.13, $CI_{95\%}$[3.64, 4.61]). Thus, people in JE did not overpredict the difference in happiness between the 0 and 80 buyers scenarios. Furthermore, and consistent with the conclusions of Hsee and Zhang's (2004) original studies, people in JE underpredicted the happiness ratings compared people in SE—the effect size for the SE comparison (i.e., [3.64, 4.61]) is larger than the JE comparison (i.e., [2.11, 3.1]) and the confidence intervals do not overlap.

Taken together, the results of Study 1 in the present paper support the theory of distinction bias and successfully replicate the results of Study 1 from Hsee and Zhang (2004). More precisely, people in JE overpredicted the differences in happiness for the quantitatively different scenarios but not for the qualitatively different scenarios. See Table S1 for a comparison of the inferential statistics between the original and replication studies.

We intended to assess the replication results by comparing effect sizes between the original and replication studies (see LeBel et al., 2019), but the original results in the paper by Hsee and

Zhang (2004) were not reported in sufficient detail to allow for such comparisons, which further points to the importance of complete statistical reporting in scientific articles (see Bakker & Wicherts, 2011; Olsen et al., 2019). Nonetheless, (in Supplemental Materials) we present effect size comparisons between the replication and original studies, where possible.

Extension results. In the present study we included an additional JE group and two additional SE groups to examine the effect of introducing an alternative reference point in the scenarios to examine the distinction bias phenomenon. In the original and replication design, the 0 buyers and 80 buyers scenarios are values that cross the reference point (qualitative difference), whereas the 80 buyers, 160 buyers, and 240 buyers scenarios are on the same side of the reference point (quantitative differences). Hsee and Zhang (2004) argued that the threshold point may lie elsewhere, such that if the threshold is at 160 buyers then the difference between 80 buyers and 240 buyers becomes qualitative and people in JE should no longer overpredict the difference in happiness for these two scenarios compared to people in SE. To test this, we included a JE extension group and two SE extension groups (80 buyers scenario and 240 buyers scenario), all of whom were told that they had expected 160 books to be sold.

The results showed that people in the JE extension group predicted statistically significant higher happiness for the 240 buyers scenario than the 80 buyers scenario, $t_{(101)} = 16.4$, $p < .001$, $d = 2.27$, $CI_{95\%}[1.76, 2.79]$. Likewise, comparing the SE extension groups showed that those in the 240 buyers scenario predicted significantly higher happiness than those in the 80 buyers scenario, $t_{(151.37)} = 10.38$, $p < .001$, $d = 1.45$, $CI_{95\%}[1.14, 1.75]$ suggesting that the difference between the two scenarios was no longer overpredicted by people in JE. However, given that the effect size for the JE comparison is larger than the effect size for the SE comparison and the confidence intervals do not overlap, it could be argued that the data are consistent with the idea that people in JE extension overpredicted happiness compared to people in SE extensions. On the other hand, although the

confidence intervals do not overlap, they are quite close (i.e., 1.76 vs 1.75; also see the slopes in Figure 3), especially compared to the confidence interval for the difference between the original SE groups in the 80 and 240 buyers scenarios when the reference point was 0, $t_{(206.58)} = 0.84$, $p = .40$, $d = 0.12$, $CI_{95\%}$ [-0.16, 0.39]. Therefore, what was previously a quantitative difference (i.e., 80 vs 240 buyers) was rendered a qualitative difference when an alternative reference point was introduced (i.e., 160 buyers), and people in JE no longer substantially overpredicted the difference in happiness between options as much as when the reference point was 0. In other words, as the theory predicted, distinction bias occurred more strongly when the difference in values were quantitative compared to when they were qualitative.

Study 2 (Words Task)

Methods

Materials, measures, and procedure. Participants were randomly allocated to one of nine groups (one JE group, four SE real-experience groups, and four SE predicted-experience groups).

JE predicted-experience group. Participants in the JE predicted-experience group were presented with four tasks and were asked to imagine they had finished all four. The four tasks were:

> Read and copy-paste a list of <u>25 negative</u> words, such as *hatred* and *loss*.
>
> Read and copy-paste a list of <u>10 negative</u> words, such as *hatred* and *loss*.
>
> Read and copy-paste a list of <u>10 positive</u> words, such as *love* and *win*.
>
> Read and copy-paste a list of <u>25 positive</u> words, such as *love* and *win*.

After presenting all four tasks together, each task description was presented in turn and participants in the JE predicted-experience group were asked to rate how happy they would feel (on the same 9-point scale as in Study 1) if they had actually performed the task.

SE real-experience group. Participants in the four SE real-experience groups were each presented with only one of the four tasks, without knowing about the other three tasks, and asked to actually perform the task. They were shown the list of 10 or 25 positive or negative words (according to the condition to which they were assigned), presented horizontally and separated by commas, and then asked to copy and paste the words into the correctly labeled textboxes presented in a column below the horizontal list of words. The words could only be copied and pasted individually. Participants could not proceed to the next page unless they entered the right keyword in the right textbox for each of those words. Therefore, for each word on the list, participants had to 1) read the word, 2) copy the word, and then 3) paste the word in the correct textbox. Given that this had to be done for each of the words, the task was longer for the longer list of words. After they finished the task, they were asked to rate how they felt on the 9-point happiness scale.

SE predicted-experience group. Participants in the four SE predicted-experience groups were also presented with only one of the four tasks, without knowing about the other three tasks, and asked to predict how they would feel after completing the task on the 9-point happiness scale.

Thus, the ratings given by the JE predicted-experience and the SE predicted-experience groups were predictions of happiness whereas the ratings given by SE real-experience groups were the experienced happiness after actually completing the task.

Results and Discussion

The descriptive statistics for the ratings on predicted happiness for each scenario partitioned by condition, for Study 2, are presented in Figure 4 and Table 2. We used Student's paired samples *t*-tests for comparisons between scenarios for the JE predicted-experience group and Welch's independent samples *t*-tests for the comparisons between the SE groups. All tests were two-tailed. According to the theory of distinction bias, compared to people in SE real-experience and SE predicted-experience, people in JE predicted-experience were expected to (1) overpredict the

difference in happiness between the scenario with 10 negative (positive) words and the scenario with 25 negative (positive) words (i.e., the quantitatively different scenarios), and (2) not overpredict the difference between the scenario with 10 negative words and the scenario with 10 positive words (i.e., the qualitatively different scenarios). The pattern of results do not support the distinction bias hypotheses.

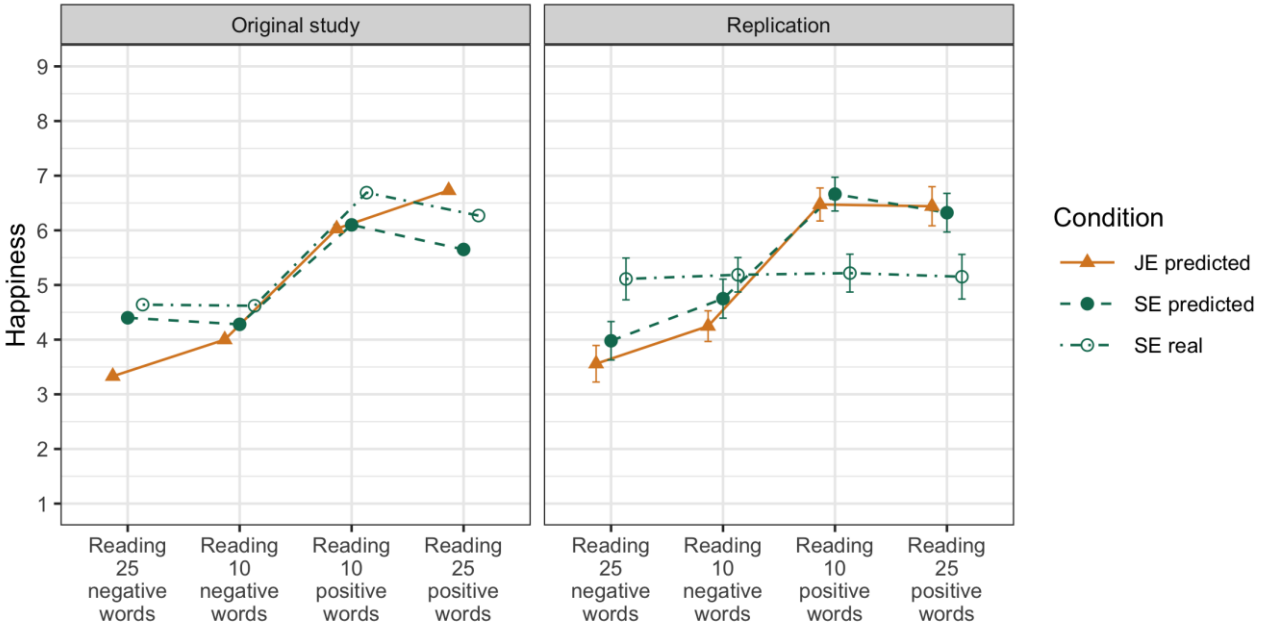

Figure 4. Word study: People in joint evaluation (JE) overpredicted the difference in happiness between copy-pasting 25 negative words and copy-pasting 10 negative words, and between copy-pasting 10 negative words and copy-pasting 10 positive words, in comparison to SE real-experience but not in comparison to SE predicted-experience. But people in JE did not overpredict the difference between copy-pasting 10 positive words and copy-pasting 25 positive words compared to either the SE real-experience or SE predicted-experience comparisons.

Table 2
Study 2 Descriptive Statistics for Original Study and Replication Study

Evaluation mode	Type of reading	Original Study		Replication Study		
		N	*M*	*N*	*M*	*SD*
JE pred.	25 neg.	360/9	3.33	93	3.56	1.64
JE pred.	10 neg.		4.00		4.25	1.38
JE pred.	10 pos.		6.03		6.47	1.49
JE pred.	25 pos.		6.73		6.44	1.76
SE pred.	25 neg.	360/9	4.40	95	3.98	1.75
SE pred.	10 neg.	360/9	4.28	92	4.75	1.75
SE pred.	10 pos.	360/9	6.10	92	6.66	1.51
SE pred.	25 pos.	360/9	5.65	93	6.32	1.74
SE real	25 neg.	360/9	4.64	90	5.11	1.85
SE real	10 neg.	360/9	4.62	91	5.19	1.53
SE real	10 pos.	360/9	6.69	92	5.22	1.70
SE real	25 pos.	360/9	6.27	86	5.15	1.93

Note. For the original study only the overall sample size was reported. SDs were not reported at all. JE pred = JE predicted-experience. SE pred = SE predicted-experience. SE real = SE real-experience.

Distinction bias replication results. When examining the quantitatively different scenarios, people in JE predicted-experience predicted statistically significant higher happiness for the scenario with 10 negative words than the scenario with 25 negative words ($t_{(92)} = 5.06$, $p < .001$, $d = 0.45$, $CI_{95\%}[0.26, 0.63]$), but they did not predict statistically significant higher happiness for the scenario with 25 positive words compared to the scenario with 10 positive words ($t_{(92)} = 0.19$, $p = 0.847$, $d = 0.02$, $CI_{95\%}[-0.18, 0.22]$). People in the different SE real-experience groups did not give significantly higher happiness ratings for the scenario with 10 negative words than the scenario with 25 negative words ($t_{(172.35)} = 0.3$, $p = .765$, $d = 0.04$, $CI_{95\%}[-0.25, 0.34]$) nor did they give significantly higher happiness ratings for the scenario with 25 positive words compared to the scenario with 10 positive words ($t_{(169.45)} = 0.24$, $p = 0.809$, $d = 0.04$, $CI_{95\%}[-0.26, 0.33]$). It could be

argued that, because the JE predicted-experience group predicted less happiness for the 25 negative words compared to the 10 negative words whereas the corresponding SE real-experience groups did not show statistically significant differences, people in JE predicted-experience overpredicted the differences in these tasks. However, the effect size confidence intervals overlap, suggesting that the data are also consistent with the idea that there was no overprediction. There was no overprediction for the quantitatively difference scenarios for positive words when comparing JE predicted-experience against SE real-experience groups.

Results of participants in the different SE predicted-experience groups follow the same pattern as for people in the JE predicted-experience group, with lower predicted happiness for the scenario with 25 negative words than the scenario with 10 negative words ($t_{(184.84)} = 3.02$, $p = 0.003$, $d = 0.44$, $CI_{95\%}[0.15, 0.73]$), but there was not significantly lower happiness for the scenario with 10 positive words compared to the scenario with 25 positive words ($t_{(180.09)} = 1.42$, $p = .157$, $d = 0.21$, $CI_{95\%}[-0.08, 0.50]$). Thus, people in JE predicted-experience did not overpredict the difference in happiness between the quantitatively different scenarios when compared to people in SE predicted-experience, particularly given the overlap of the effect size confidence intervals.

In sum, the results (at best) support the idea that people in JE predicted-experience did exhibit distinction bias when compared to people in the SE real-experience groups for the negative words; but we found no support for the idea that people in JE predicted-experience exhibited distinction bias when compared to people in the SE predicted-experience groups or when compared to the people in the SE real-experience groups for the positive words.

When examining the qualitatively different scenarios (i.e., 10 negative words vs 10 positive words), we see that people in JE predicted-experience predicted significantly higher happiness for the scenario with 10 positive words, $t_{(92)} = 10.24$, $p < .001$, $d = 1.55$, $CI_{95\%}[1.11, 2]$. However, people in the SE real-experience groups did not give higher happiness ratings for the scenario with

10 positive words compared to the scenario with 10 negative words, $t_{(179.58)} = 0.13$, $p = .898$, d = 0.02, $CI_{95\%}$[-0.27, 0.31]. On the other hand, people in the SE predicted-experience groups predicted significantly greater happiness for the scenario with 10 positive words than the scenario with 10 negative words, $t_{(178.43)} = 7.94$, $p < .001$, $d = 1.17$, $CI_{95\%}$[0.86, 1.49]. Thus, people in JE predicted-experience overpredicted the difference in happiness for the qualitatively different scenarios when compared to people in the SE real-experience, which is inconsistent with the theory of distinction bias—note that the confidence intervals do not overlap—but they did not overpredict the difference when compared to people in SE predicted-experience, which is consistent with the theory of distinction bias—note that the confidence intervals overlap.

The results of Study 2 in the present paper are inconclusive and, in general, do not replicate the focal results of Study 2 from Hsee and Zhang (2004). The theoretical predictions were supported only in one out of four quantitative comparisons (overprediction in JE predicted-experience of 25 vs. 10 negative words in comparison to SE real-experience) and in one of two qualitative comparisons (no overprediction in JE predicted-experience of 10 negative vs. 10 positive words in comparison to SE predicted-experience). Note that for the quantitatively different scenarios, the distinction bias hypothesis was supported for the comparison between JE predicted-experience and SE real-experience, whereas for the qualitatively different scenarios the hypothesis was supported for the comparison between JE predicted-experience and SE predicted-experience.

Figure 4 illustrates clearly how the replication results deviate from the original study's results and comparisons of inferential statistics are presented in Table S1. One clear deviation between the results of the original and replication studies is that the two different SE (experienced and predicted) conditions followed the same pattern as each other in the original study (significant differences between the qualitatively different scenarios but not between the quantitatively different scenarios), but they diverged in the replication study. In the replication, the SE real-experience

groups gave similar happiness ratings regardless of the scenario whereas the SE predicted-experience groups gave successively higher happiness ratings from 25 negative words to 10 negative words and then to 10 positive words, but there was a nonsignificant difference between 10 positive words and 25 positive words. One explanation may be that the manipulation of having to copy-paste the words was not strong enough to influence happiness ratings; for example, because participants in the SE-experience condition may have copy-pasted the words without processing their meaning, thus remaining insensitive to their valence. However, we consider that an unlikely explanation given that our adjustments to this study—that of requiring participants to paste each word in the correct textbox—were made to increase the likelihood that participants actually read the words. Moreover, this would not explain the deviations between the original and replication study with regard to the comparisons of JE predicted-experience and SE predicted-experience conditions, where (in the replication study) overpredictions of quantitative judgments were absent. We provide potential explanations for this in the general discussion.

General Discussion

We conducted very close replications of Studies 1 and 2 from Hsee and Zhang (2004). The original findings provided evidence for distinction bias which proposes that a difference between evaluation modes (i.e., joint evaluation and single evaluation, JE and SE respectively) is a key mechanism for affective mispredictions. Our data set, consisting of approximately two times more participants per cell than the original studies, provided mixed support for distinction bias. Study 1 fully replicated the original study's results whereas Study 2 did not replicate (i.e., find statistically significant effects in the same direction as) most of the original study's findings.

Study 1 provided evidence supporting distinction bias. For quantitatively different options, the replication study detected a signal in the same direction as the original study for people in JE and detected no signal for people in SE. Critically, in line with the predictions of distinction bias,

people in JE overpredicted happiness differences for quantitatively different options when compared to people in SE. For qualitatively different options, the replication study detected signals in the same direction as the original study for people in JE and people in SE. And, in line with the predictions of distinction bias, people in JE did not overpredict (but rather underpredicted) happiness differences for qualitatively different options when compared to people in SE. Taken together, the results of Study 1 fully replicated the original findings.

Study 1 included an extension that tested another hypothesis put forward by Hsee and Zhang (2004). Namely, the distinction bias explanation for affective mispredictions suggests that if the reference point is changed, such that what were previously considered quantitatively different options become qualitatively different because they cross over the reference point, then people in JE should come to overpredict the differences between these options. We found that when the reference point was changed (i.e., participants were told that the expectation was that they would sell 160 books), the difference in happiness levels between options that now crossed the reference point (i.e., selling 80 or 240 books) for people in SE became statistically significant (for reference point of 160 books, SE: $CI_{95\%}$[1.14, 1.75]; for reference point of 0 books, SE: $CI_{95\%}$[-0.16, 0.39]).

In contrast to Study 1, the replication results of Study 2 were not in line with the original study and failed to support the distinction bias hypotheses in three ways. First, people in the SE real-experience groups did not differ in reported levels of happiness across the (quantitatively or qualitatively) different scenarios—distinction bias predicts that there should be a difference in happiness levels when the scenarios are qualitatively different. Second, people in JE predicted differences between the qualitatively different options (i.e., 10 negative words vs 10 positive words). This goes further against distinction bias, given that people in JE overpredicted (when, according to distinction bias, they should not have overpredicted) the difference between the qualitatively different options relative to people in SE real-experience. Third, differences in

happiness levels between the different scenarios for people in the SE predicted-experience groups followed the same pattern as for people in the JE predicted-experience group, such that people in JE predicted-experience made no overpredictions relative to people in SE predicted-experience.

Why do we find mixed results between the two replication studies? One explanation for the deviations could be the real-world relevance of the scenarios. Selling books that one has written (Study 1) has a more obvious real-world relevance for one's happiness than merely reading words (Study 2). The selling books scenario can be considered as a hypothetical personal achievement as well as a situation of hypothetical monetary gain. We believe this is more concrete with higher personal relevance than imagining or really reading 10 or 25 words of different valence. Therefore, we believe that reading 10 or 25 words is unlikely to influence how happy people are. Indeed, we find no evidence that they do, given the results of the SE real-experience groups. In our opinion, therefore, the design of Study 2 was not adequate for testing the distinction bias hypothesis. Perhaps a better implementation would be to use images, or some other more meaningful task, such as writing about a sad/happy experience from the past, to induce (un)happiness. Essentially, we believe that Study 1 provides a more convincing test of the distinction bias hypothesis than Study 2. Taken together, because we believe that (a) Study 1 is the better test of the theory than Study 2, (b) our findings for Study 1 replicate the original results, and (c) the results of our extension in Study 1 further support the theory behind distinction bias, we believe that our findings strengthen the evidence for the theory.

Yet, this raises the question why the original Study 2 found such clear results in support of distinction bias using the very same materials. The inconsistency may be due to differences in the replication procedure (see Rosenthal, 1990). For example, the present study was conducted online with an MTurk sample. However, online samples from Amazon Mechanical Turk have been shown to provide reliable data (Buhrmester et al., 2016; Coppock, 2019; Coppock et al., 2018; Zwaan et

al., 2018). And, as noted earlier, evidence suggests that little variability in results between studies can be attributed to whether the study was run in the lab or online (e.g., Klein et al., 2018). On the other hand, it could also be the case that the results from the original paper were due to other idiosyncratic aspects of the design (e.g., population sampled from or research site) rather than the phenomenon under study.

Nevertheless, it might be argued that participants engaged in an online study may attend less to the task at hand, wanting to complete the study as quickly as possible, when compared to participants in the lab. Using this logic, the SE-experienced group may have been copy-pasting the words without processing their meaning, thus being insensitive to the valence. Notwithstanding that this argument could equally be applied to students coming into the lab, who would want to complete the study as quickly as possible as well, we adjusted the reading task to a read and copy-paste task in order to give some assurance that participants indeed read the words. To examine this, we conducted an MTurk study with 201 participants randomly allocated to copy and paste 10 or 25 positive or negative words into textboxes, the exact same task as in Study 2 (SE-experienced). We asked participants, with yes/no response options, whether they read the words and whether they processed them: 89% reported having read the words and 62% that they processed the meaning of the words (see details in the supplementary materials). While this provides support for our interpretation that participants read the words, we cannot completely rule out the possibility that *some* participants may not have processed the meaning of the words they were reading and copy-pasting, though a similar issue exists with the target article's task. Nevertheless, even if inattention to the words explained why the distinction bias effect was not observed when comparing the JE-predicted and SE-experienced groups, it would not explain why we did not observe a distinction bias effect when comparing JE-predicted with SE-predicted. Thus, the explanation that MTurk

workers did not attend to the task or completed the task less attentively would not explain the entirety of the failure of Study 2 to find support for distinction bias.

In sum, we successfully replicated Study 1, from Hsee and Zhang (2004), but not Study 2. The findings suggest that distinction bias may occur only for decisions or events that have some meaning for people, but perhaps not for insubstantial priming tasks such as reading words. Future research on distinction bias should thus use more meaningful tasks and decisions and perhaps aim at identifying, more conclusively, whether meaningfulness of a task or event moderates the occurrence of distinction bias, or whether the discrepancy between the results of our replication and those in the original Study 2 is due to some hidden factor or an adjustment made in our replication. Another avenue for future research is with regards to generalizability. Although we have discussed the results and put forth our conclusions in general terms, we acknowledge that the present results (and the results of the original studies) cannot safely be generalized beyond the single task and single set of stimuli that were used. To be able to generalize across tasks and stimuli, future research should randomly sample a selection of tasks and stimuli and model them as random factors in the statistical models (see Yarkoni, 2019, for a full treatment of generalizability).

References

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678. https://doi.org/10/fdz26x

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & Van't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Brooks, M. E., Guidroz, A. M., & Chakrabarti, M. (2009). Distinction bias in applicant reactions to using diversity information in selection. *International Journal of Selection and Assessment*, *17*(4), 377–390. https://doi.org/10.1111/j.1468-2389.2009.00480.x

Buhrmester, M., Kwang, T., & Gosling, S. D. (2016). *Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality data?* https://doi.org/10/gghxfn

Camerer, C. F., Dreber, A., Forsell, E., Ho, T., Huber, J., Kirchler, M., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*, 1433–1436. https://doi.org/10.1126/science.aaf0918

Coles, N., Tiokhin, L., Scheel, A., Isager, P. M., & Lakens, D. (2018). The costs and benefits of replication studies. *Behavioral and Brain Sciences*, *41*, 124. https://doi.org/10.1017/S0140525X18000596

Coppock, A. (2019). Generalizing from survey experiments conducted on Mechanical Turk: A replication approach. *Political Science Research and Methods*, *7*(3), 613–628. https://doi.org/10/gfrbm7

Coppock, A., Leeper, T. J., & Mullinix, K. J. (2018). Generalizability of heterogeneous treatment effect estimates across samples. *Proceedings of the National Academy of Sciences*, *115*(49), 12441–12446. https://doi.org/10/ggf5zm

Dunn, E. W., Wilson, T. D., & Gilbert, D. T. (2003). Location, location, location: The misprediction of satisfaction in housing lotteries. *Personality and Social Psychology Bulletin*, *29*(11), 1421–1432. https://doi.org/10.1177/0146167203256867

Dunn, T. L., Koehler, D. J., & Risko, E. F. (2017). Evaluating effort: Influences of evaluation Mode on judgments of task-specific efforts. *Journal of Behavioral Decision Making*, *30*(4), 869–888. https://doi.org/10.1002/bdm.2018

Gilbert, D. T., Gill, M. J., & Wilson, T. D. (2002). The future is now: Temporal correction in affective forecasting. *Organizational Behavior and Human Decision Processes*, *88*(1), 430–444. https://doi.org/10.1006/obhd.2001.2982

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., … Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*, 546–573. https://doi.org/10.1177/1745691616652873

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin, 125*(5), 576. https://doi.org/10/dm3fvc

Hsee, C. K., & Zhang, J. (2004). Distinction bias: Misprediction and mischoice due to joint evaluation. *Journal of Personality and Social Psychology*, *86*(5), 680. https://doi.org/10/dhnrdb

Hsee, C. K., & Zhang, J. (2010). General evaluability theory. *Perspectives on Psychological Science*, *5*(4), 343–355. https://doi.org/10.1177/1745691610374586

Hsee, C. K., Zhang, J., Wang, L., & Zhang, S. (2013). Magnitude, time, and risk differ similarly between joint and single evaluations. *Journal of Consumer Research*, *40*(1), 172–184. https://doi.org/10.1086/669484

Isager, P. M. (2019). Quantifying Replication Value: A formula-based approach to study selection in replication research. *Open Science 2019, Trier, Germany*.

Kahneman, D. (2000). Experienced utility and objective happiness: A moment-based approach. In *Choices, values, and frames* (Vol. 1, pp. 187–208). Cambridge University Press.

Kahneman, D., & Snell, J. (1992). Predicting a changing taste: Do people know what they will like? *Journal of Behavioral Decision Making*, *5*(3), 187–200. https://doi.org/10.1002/bdm.3960050304

Klein, R. A., Ratliff, K. A., Vianello, M., Adams Jr., R. B., Bahník, Š., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*, 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr., R. B., Alper, S., … Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490. https://doi.org/10.1177/2515245918810225

Kleinmuntz, D. N., & Schkade, D. A. (1993). Information displays and decision processes. *Psychological Science*, *4*(4), 221–227. https://doi.org/10.1111/j.1467-9280.1993.tb00265.x

Kurtz, J. L., Wilson, T. D., & Gilbert, D. T. (2007). Quantity versus uncertainty: When winning one prize is better than winning two. *Journal of Experimental Social Psychology*, *43*(6), 979–985. https://doi.org/10.1016/j.jesp.2006.10.020

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10/f96zbh

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

LeBel, Etienne Philippe, Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A brief guide to evaluate replications. *Meta-Psychology*, *3*. https://doi.org/10.15626/mp.2018.843

Markman, K. D., & Hirt, E. R. (2002). Social prediction and the "allegiance bias." *Social Cognition*, *20*(1), 58–86. https://doi.org/10.1521/soco.20.1.58.20943

Morewedge, C. K., Gilbert, D. T., Myrseth, K. O. R., Kassam, K. S., & Wilson, T. D. (2010). Consuming experience: Why affective forecasters overestimate comparative value. *Journal of Experimental Social Psychology*, *46*(6), 986–992. https://doi.org/10.1016/j.jesp.2010.07.010

Novemsky, N., & Ratner, R. K. (2003). The time course and impact of consumers' erroneous beliefs about hedonic contrast effects. *Journal of Consumer Research*, *29*(4), 507–516. https://doi.org/10.1086/346246

O'Brien, E., & Roney, E. (2017). Worth the wait? Leisure can be just as enjoyable with work left undone. *Psychological Science*, *28*(7), 1000–1015. https://doi.org/10.1177/0956797617701749

Olsen, J., Mosen, J., Voracek, M., & Kirchler, E. (2019). Research practices and statistical reporting quality in 250 economic psychology master's theses: A meta-research investigation. *Royal Society Open Science*, *6*(12), 190738. https://doi.org/10/ggfpr4

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

Ratner, R. K., Kahn, B. E., & Kahneman, D. (1999). Choosing less-preferred experiences for the sake of variety. *Journal of Consumer Research*, *26*(1), 1–15. https://doi.org/10.1086/209547

Rosenthal, R. (1990). Replication in behavioral research. *Journal of Social Behavior and Personality*, *5*(4), 1.

Schkade, D. A., & Kahneman, D. (1998). Does living in California make people happy? A focusing illusion in judgments of life satisfaction. *Psychological Science*, *9*(5), 340–346. https://doi.org/10.1111/1467-9280.00066

Scott, J. (2000). Rational choice theory. In G. Browning, A. Halcli, and F. Webster, *Understanding contemporary society: Theories of the present*. 126-138. Sage Publications https://doi.org/10.4135/9781446218310.n9

Shen, L., Hsee, C. K., Wu, Q., & Tsai, C. I. (2012). Overpredicting and underprofiting in pricing decisions. *Journal of Behavioral Decision Making*, *25*(5), 512–521. https://doi.org/10/b8h26j

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

Wilson, T. D., & Gilbert, D. T. (2003). Affective forecasting. *Advances in Experimental Social Psychology*, *35*(35), 345–411. https://doi.org/10.1016/s0065-2601(03)01006-2

Yarkoni, T. (2019). *The generalizability crisis*. https://doi.org/10/ggdf7h

Ziano, I., Wang, Y. J., Sany, S. S., Feldman, G., Ngai, L. H., Lau, Y. K., Bhattal, I. K., Keung, P. S., Wong, Y. T., & Tong, W. Z. (2019). Perceived morality of direct versus indirect harm: Replications of the preference for indirect harm effect. *PsyArXiv*. https://doi.org/10.31234/osf.io/bs7jf

Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (2018). Participant nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin & Review*, *25*(5), 1968–1972. https://doi.org/10/gff4sz

Effect Size Comparisons: Original vs Replication

To assess replication results, LeBel et al. (2019) proposed three criteria: (i) whether a signal was detected in the replication study (i.e. a statistically significant result in the same direction of the original study), (ii) whether the effect size of the replication study is consistent with the effect size in the original study (i.e., whether the 95% confidence interval of the replication study contains the point estimate of the original study), and (iii) the relative precision of the effect size estimates in the replication and the original studies (i.e., the relative width of the 95% confidence intervals). However, the original paper by Hsee and Zhang (2004) did not report all statistics necessary to calculate confidence intervals and only some results were reported such that we could calculate effect sizes (we contacted the original authors but they were unable to provide the original data, given how long ago the studies were carried out). Thus, to assess the replication results, we assessed whether the replication study detected a signal and, where possible, whether it was consistent with the effect size of the original study.

Study 1

For quantitatively different options, in both the original study by Hsee and Zhang (2004) and the present replication study, people in the JE group predicted significantly greater happiness for the 160 vs 80 buyers scenarios and significantly greater happiness for the 240 vs 160 buyers scenarios. For the corresponding SE groups, there were nonsignificant differences in both the original and replication studies between the 80 and 160 buyers scenarios, and also between the 160 and 240 buyers scenarios. Thus, the replication detected a signal in the same direction as the original study for the quantitatively different scenarios for the JE group and, also in line with the original study, nonsignificant differences (i.e., did not detect a signal) between the SE groups. In

other words, the results of the replication study were in line with the results of the original study for the quantitatively different options.

Regarding the qualitatively different options (i.e. 0 and 80 buyers scenarios), the original and replication JE groups predicted that they would be happier in the 80 buyer scenario, $d_{original}$ = 1.17 and $d_{replication}$ = 2.60 ($CI_{95\%}$[2.11, 3.10]). The difference between the 0 and 80 buyers scenarios for the SE groups in the original and replication studies were also both statistically significant, $d_{original}$ = 3.26 and $d_{replication}$ = 4.13 ($CI_{95\%}$[3.64, 4.61]). Thus, for the qualitatively different options, the replication results detected a signal in the same direction as the original study's results; though the confidence intervals of the replication study did not include the effect size point estimates from the original (the effect sizes in the replication were larger). We can conclude that the replication results aligned with the original also for the qualitatively different scenarios.

Study 2

For the quantitatively different options, we will examine the scenarios with the negative words separately from the scenarios with the positive words. People in JE predicted-experience predicted statistically significant greater unhappiness for the 25 negative words scenario than the 10 negative words scenario in both studies, $d_{original}$ = 0.60 and $d_{replication}$ = 0.45 ($CI_{95\%}$[0.26, 0.63]). These results from the replication study were signal consistent with the original study such that the replication detected a signal in the same direction and had an effect size confidence interval containing the point estimate from the original study. For people in SE real-experience groups, there were nonsignificant differences in unhappiness ratings between the 25 negative words than the 10 negative words in both studies, which aligned with the results of the original

study. On the other hand, for the SE predicted-experience groups in the original study the difference between 25 negative words and 10 negative words was not statistically significant, whereas in the replication study this difference was significant.

Regarding positive words, people in JE predicted-experience predicted statistically significant lower happiness for the 10 positive words than the 25 positive words in the original study, $d_{original} = 0.75$, but this difference was not statistically significant in the replication study, $d_{replication} = 0.02$, $CI_{95\%}[-0.18, 0.22]$). For JE predicted-experience, in contrast to the original study, the replication did not detect a signal. In the original study, there was not a statistically significant difference in happiness ratings between 10 and 25 positive words for either the SE real-experience or the SE predicted-experience groups. In line with the original study's results, these differences were not statistically significant for the SE real-experience or SE predicted-experience groups in the replication study.

Now we examine the qualitatively different tasks (10 negative vs 10 positive words). For JE predicted-experience, the replication study detected a signal in the same direction as the original study and the confidence interval shows that the replication effect size was larger than the effect size point estimate in the original study: $d_{original} = 0.91$, $d_{replication} = 1.55$, $CI_{95\%}[1.11, 2]$. For SE real-experience, the replication study did not detect a signal and the confidence interval did not contain the effect size point estimate from the original study: $d_{original} = 1.20$, $d_{replication} = 0.02$, $CI_{95\%}[-0.27, 0.31]$. Thus, in this case, the replication study was inconsistent with the original study. For SE predicted-experience, the replication detected a signal in the same direction as the original paper (though it should be noted that the inferential statistics for the SE predicted-experience groups were not reported in the original, only a qualitative statement was made about results of these groups being similar to the results of the SE real-experience groups).

Thus, for the qualitatively different tasks, the replication study was consistent with the original study for the JE predicted-experience group and the SE predicted-experience groups but not for the SE real-experience groups.

Certainty

In study 1, to get a measure of certainty about their judgments for quantitative differences, we asked participants in the JE group which scenario would make them happiest (80, 160, or 240 people bought the book). This was after they provided happiness ratings for each situation and only had to indicate which of the three situation would make them happiest. Next, they had to rate how certain they were about this judgment; for certainty about judgments for qualitative differences, we asked whether they would feel happier if no one had bought the book or if 80 people had bought the book, and then to rate how certain they were about this judgment.

In study 2, for quantitative differences, we asked the JE group to predict whether the task with 10 negative words or 25 negative words would make them happier and then to rate how certain they were about this judgment (and we did the same for the positive words); for qualitative differences, we asked the JE group to predict whether the task with 10 positive words or the task with 10 negative words would make them happier and then to rate their certainty about this judgment. We expected that, for both studies, people in JE would be more certain about their judgments for qualitative differences than about their judgments for quantitative differences.

Study 1

We asked participants in the JE group to compare the situations where 80, 160, and 240 people bought the book and indicate which would make them feel happiest (or all equally),

followed by a question asking them to rate how certain they were about that judgment (from *1 = not at all certain*, to *9 = extremely certain*). This rating provided an indication of how certain they were regarding their judgments about the difference between quantitatively different outcomes. We also asked them to compare "no one has bought your book" with "80 people have bought your book" and to choose which would make them feel happier (no one, 80 people, or equally happy), followed by the same certainty rating regarding their judgment. This rating provided an indication of how certain they were regarding their judgments about the difference between qualitatively different outcomes. These two certainty ratings were presented to participants in randomized order.

Participants in the JE extension group were also asked to indicate whether they would be happier if 80 or 240 people bought the book followed by their certainty judgment (now about the qualitative difference since 80 and 240 fall on either side of the threshold), and then whether they would feel happier if no one bought the book or if 80 people bought it followed by a rating of how certain they were about this judgment (regarding the quantitative difference, since 0 and 80 were assumed to fall on the same side of the 160 threshold). The certainty ratings were given on the same scale as described for the JE group previously.

People in JE did not significantly differ in their certainty for judging the quantitatively different scenarios (M = 8.46, SD = 0.89) compared to the qualitatively different scenarios (M = 8.51, SD = 0.88), $t_{(102)} = 0.66$, $p = .51$, $d = 0.07$, $CI_{95\%}$ [-0.13, 0.26].

People in the JE extension group gave significantly lower certainty ratings for their judgments about quantitative differences (whether 0 buyers or 80 buyers would make them happier; M = 8.40, SD = 1.15) than their judgments about qualitative differences (whether 80 or

240 buyers would make them happier; $M = 8.63$, $SD = 1.01$), $t_{(101)} = 2.59$, $p = .011$, $d = 0.21$, $CI_{95\%}$ [0.05, 0.37].

Study 2

Participants in JE were asked to compare the 10 negative words task with the 10 positive words task and choose in which scenario they would experience more happiness or whether they would be equally happy, followed by a rating for how certain they were about this judgment (about qualitative differences). JE participants were also asked to compare the 25 negative words task with the 10 negative words task, and the 10 positive words task with the 25 positive words task, and indicate which of these two would make them happiest (or whether they would be equally happy), followed by a rating of how certain they were for each of these judgments. The certainty ratings were given on the same 9-point scale as from study 1.

For quantitative differences, people in JE were asked to choose both whether the 10 positive or 25 positive words task, and whether the 10 negative or 25 negative words task, would make them happiest and then to rate how certain they were in both of these judgments. For qualitative differences, people in JE were asked to choose whether the 10 negative words or the 10 positive words task would make them happiest and then to rate their certainty about this judgment.

People in JE had less certainty for judging the quantitatively different scenarios ($M = 7.15$, $SD = 1.61$) than the qualitatively different scenarios ($M = 7.73$, $SD = 1.32$), $t_{(92)} = 4.32$, $p < .001$, $d = 0.39$, $CI_{95\%}$ [0.20, 0.57] when comparing certainty ratings for the quantitative difference in negative words (10 negative or 25 negative words task) with the qualitative certainty rating. However, this effect was not observed when comparing the ratings for the quantitative difference

in positive words (10 positive or 25 positive words task; M = 7.71, SD = 1.29) with the qualitative certainty rating (M = 7.73, SD = 1.32), $t_{(92)}$ = 0.18, p = .856, d = 0.02, $CI_{95\%}$ [-0.16, 0.20].

Exploration of Study Exclusion Criteria

We explored whether excluding participants who self-reported low English proficiency as well as not being serious about filling in the survey.[1] The total sample size was N = 827. When looking at the two exclusion criteria independently, 41 individuals self-reported English proficiency below the scale maximum and 9 individuals self-reported not filling in the survey seriously. When combining both criteria, 46 individuals failed to meet one or both of the inclusion criteria. We compared the main results without any exclusions against the results after combined exclusions. As Table S6 reveals, there were no qualitative difference between the two analyses.

Study reexamining "SE experience" copy-paste task in Study 2

We ran an additional study with 201 participants to examine our adjusted copy-paste task.
Materials: https://osf.io/h473w/?view_only=42abb1c032f444d19f2dba0d834b5106
Data: https://osf.io/6eqh3/?view_only=42abb1c032f444d19f2dba0d834b5106
JAMOVI analyses code and findings:
https://osf.io/6sku4/?view_only=42abb1c032f444d19f2dba0d834b5106

[1] The preregistration states that the self-report on English proficiency would be on a 5-point scale and that anybody self-reporting below 5 would be excluded, but it was actually on a 7-point scale so we used 7 as a criterion. The preregistration proposed to examine open responses and code for hypothesis guessing, but we did not examine this criterion for exclusions.

Supplemental Tables

Table S1
Inferential Statistics from Original and Replication Studies

Comparison	Condition	Original study	Replication study
Study 1			
0 vs. 80 books	JE	t(49) = 8.20, p < .001	t(102) = -21.66, p < .001, d = -2.6, 95% CI [-3.1, -2.11]
	SE	t(98) = 16.15, p < .001	t(189.04) = -29.65, p < .001, d = -4.13, 95% CI [-4.61, -3.64]
80 vs. 160 books	JE	t(49) > 3, p < .001	t(102) = -11.08, p < .001, d = -0.77, 95% CI [-0.92, -0.61]
	SE	t(97) < 1, ns	t(198.45) = 0.62, p = .536, d = 0.09, 95% CI [-0.19, 0.36]
160 vs. 240 books	JE	t(49) > 3, p < .001	t(102) = -6.01, p < .001, d = -0.6, 95% CI [-0.82, -0.39]
	SE	t(97) < 1, ns	t(201.4) = 0.18, p = .854, d = 0.03, 95% CI [-0.25, 0.3]
80 vs. 240 books (extension)	JE		t(101) = -16.4, p < .001, d = -2.27, 95% CI [-2.79, -1.76]
	SE		t(151.37) = -10.38, p < .001, d = -1.45, 95% CI [-1.75, -1.14]
Study 2			
25 neg. vs. 10 neg. words	JE pred-exp.	t(39) = 3.74, p < .001	t(92) = -5.06, p < .001, d = -0.45, 95% CI [-0.63, -0.26]
	SE real-exp.	t > 1, ns	t(172.35) = -0.3, p = .765, d = -0.04, 95% CI [-0.34, 0.25]
	SE pred-exp.	not reported	t(184.84) = -3.02, p = .003, d = -0.44, 95% CI [-0.73, -0.15]
10 neg. vs. 10 pos. words	JE pred-exp.	t(39) = 5.67, p < .001	t(92) = -10.24, p < .001, d = -1.55, 95% CI [-2, -1.11]
	SE real-exp.	t(83) = 5.47, p < .001	t(179.58) = -0.13, p = .898, d = -0.02, 95% CI [-0.31, 0.27]
	SE pred-exp.	not reported	t(178.43) = -7.94, p < .001, d = -1.17, 95% CI [-1.49, -0.86]
10 pos. vs. 25 pos. words	JE pred-exp.	t(39) = 4.71, p < .001	t(92) = 0.19, p = .847, d = 0.02, 95% CI [-0.18, 0.22]
	SE real-exp.	t(80) = 1.23, ns	t(169.45) = 0.24, p = .809, d = 0.04, 95% CI [-0.26, 0.33]

	SE pred-exp.	not reported	t(180.09) = 1.42, p = .157, d = 0.21, 95% CI [-0.08, 0.5]

Note. pred-exp = predicted-experience. Real-exp = real-experience.

Table S2
Classification of the Replication Closeness Based on LeBel et al. (2018)

Design facet	Study 1	Study 2
Effect, Hypothesis	Same	Same
IV construct	Same	Same
DV construct	Same	Same
IV operationalization	Same	Same
DV operationalization	Same	Same
IV stimuli	Same	Same
DV stimuli	Same	Same
Population (e.g., age)	Unknown	Unknown
Procedural details	Different	Different
Physical settings	Different	Different
Replication classification	Very close replication	Very close replication

Table S3
Comparison of Hypotheses

Hypothesis	Original study	Replication study
1 (in general)	"If [the two choice options] x_1 and x_2 are merely *quantitatively different*, that is, if x_1 and x_2 have the same valence or are on the same side of a reference point, then people in JE are likely to overpredict the experiential difference these values will create in SE."	"When people in JE are asked to predict the happiness between two people presented with two different options, they overpredict the difference when the different options are quantitatively different…
2 (in general)	"If x_1 and x_2 are *qualitatively different*, that is, if x_1 and x_2 involve different valences or one is above the reference point and one below, then people in JE are not likely to overpredict the experiential difference these values will create in SE."	… but not when they are qualitatively different."
1 (for Study 1)	"People in JE would overpredict the difference in happiness between the 80-buyer, the 160-buyer, and the 240-buyer scenarios."	"Hsee and Zhang hypothesized that participants in JE would overpredict (compared to those in SE) the difference in happiness between the 80, 160, and 240 buyers scenarios, …
2 (for Study 1)	"People in JE would not (at least were less likely to) overpredict the difference in happiness between the no-buyer and the 80-buyer scnearios."	… but that they would be less likely to make such an overprediction between the no-buyers and the 80-buyers scenarios."
1 (for Study 2)	"We predicted that people in JE would pverpredict the difference in experience between these two tasks [(i.e., 25-negative-word task vs. 10-negative-word-task)]. By the same token, we predicted that people in JE would also overpredict the experiential difference between the 25-positive-word task and the 10-positive-word task."	"They hypothesised that people in JE would overpredict the differences in happiness between 25 and 10 negative words and between 10 and 25 positive words, …
2 (for Study 2)	"We predicted that people in JE were unlikely to overpredict the experiential difference between these two tasks [(i.e., 10-negative-word task vs. 10-positive-word-task)]."	… but not for the difference between 10 negative words and 10 positive words."

Table S4
Comparison of Recruited Samples in Original and Replication Studies.

	Original study	Replication study
Sample size	Study 1: *N* = 249 Study 2: *N* = 360	Study 1 & 2: *N* = 824
Country	United States	United States
Gender	Not reported	373 males, 451 females
Mean age	Not reported	40.2 (SD = 12)
Data collection	On campus	Online (MTurk)
Population	US Students	MTurk workers
Compensation	Not reported	Nominal payment

Table S5
Deviations from Preregistration

Domain	Preregistered	Changes, if applicable
Power analysis	For each comparison (e.g., Study 1 SE condition no-buyer vs. 80-buyer scenario) the observed effect size (e.g., Cohen's d = 1.17) was used for an apriori power analysis with $1-\beta = .95$ and $\alpha = .05$. Per study, the resulting *N*s were then added to a sum.	The original preregistered power analysis was suboptimal. First, the resulting *N*s of all power analyses ideally shouldn't have been added to a sum. Instead, it would have been better to decide on the smallest effect size of interest or downward correct the smallest observed predicted effect and calculate one power analysis for the used t-test based on this target effect size. The resulting large sample resulted from the fact that one of the power analyses was calculated for an effect that was not significant (but also predicted as such) and therefore very small. Second, some effect sizes can actually not be computed based on the reported results in the original paper. The approximation of these effects is therefore suboptimal. Instead of reporting this power analysis in the manuscript, we opted for providing information about the power our final sample sizes had for various effect sizes.
Main hypotheses tests	Independent sample t-test for differences between SE conditions Dependent sample t-test for differences between JE conditions	N/A
Additional hypotheses tests	"On top of the t-test result, equivalence testing would be used to analysis the results of the two experiments. As the t-test results of JE and SE are obtained with different t-tests (Dependent sample t-test for JE and independent sample t-test for SE), the results are no comparable. In the replication study, t-test and equivalence testing will be used. Based on the t-test results, effect sizes and confidence intervals (CIs) will be calculated and compared. Analyses method of CIs would be based on Lakens, Scheel and Isager (2018)'s article on equivalence testing. On the one hand, if the prediction of JE and SE are different (i.e. if JE overpredict or underpredict the SE), the CI would not overlap. On the other hand, if JE does not overpredict or underpredicts SE, the CIs would overlap and the effect size of the weaker effect is within the CI of the stronger effect."	We refrain from the term "equivalence testing" in the manuscript. We do compare different confidence intervals of effect sizes to evaluate whether t-test results differ from each other. However, this is not strictly "equivalence testing", which is understood as comparing a test result against zero as well as a predefined smallest effect size of interest.

Table S6
Comparison of Full vs. After-Exclusion Sample Analyses

Comparison	Condition	Full sample (N = 827)	After exclusions (N = 781)
Study 1			
0 vs. 80 books	JE	t(102) = -21.66, p < .001, d = -2.6, 95% CI [-3.1, -2.11]	t(98) = -22.35, p < .001, d = -2.73, 95% CI [-3.26, -2.21]
	SE	t(189.04) = -29.65, p < .001, d = -4.13, 95% CI [-4.61, -3.64]	t(182.08) = -30.73, < .001, d = -4.36, 95% CI [-4.88, -3.84]
80 vs. 160 books	JE	t(102) = -11.08, p < .001, d = -0.77, 95% CI [-0.92, -0.61]	t(98) = -11.58, < .001, d = -0.77, 95% CI [-0.92, -0.62]
	SE	t(198.45) = 0.62, p = .536, d = 0.09, 95% CI [-0.19, 0.36]	t(180.43) = 0.95, p = .346, d = 0.14, 95% CI [-0.15, 0.43]
160 vs. 240 books	JE	t(102) = -6.01, p < .001, d = -0.6, 95% CI [-0.82, -0.39]	t(98) = -6, p < .001, d = -0.64, 95% CI [-0.87, -0.41]
	SE	t(201.4) = 0.18, p = .854, d = 0.03, 95% CI [-0.25, 0.3]	t(185.85) = 0.06, p = .955, d = 0.01, 95% CI [-0.28, 0.29]
80 vs. 240 books (extension)	JE	t(101) = -16.4, p < .001, d = -2.27, 95% CI [-2.79, -1.76]	t(93) = -15.68, p < .001, d = -2.26, 95% CI [-2.79, -1.72]
	SE	t(151.37) = -10.38, p < .001, d = -1.45, 95% CI [-1.75, -1.14]	t(141.73) = -10.59, p < .001, d = -1.49, 95% CI [-1.8, -1.17]
Study 2			
25 neg. vs. 10 neg. words	JE pred-exp.	t(92) = -5.06, p < .001, d = -0.45, 95% CI [-0.63, -0.26]	t(88) = -4.92, p < .001, d = -0.47, 95% CI [-0.66, -0.27]
	SE real-exp.	t(172.35) = -0.3, p = .765, d = -0.04, 95% CI [-0.34, 0.25]	t(154.62) = -0.27, p = .788, d = -0.04, 95% CI [-0.35, 0.26]
	SE pred-exp.	t(184.84) = -3.02, p = .003, d = -0.44, 95% CI [-0.73, -0.15]	t(171.89) = -2.75, p = .007, d = -0.42, 95% CI [-0.72, -0.11]
10 neg. vs. 10 pos. words	JE pred-exp.	t(92) = -10.24, p < .001, d = -1.55, 95% CI [-2, -1.11]	t(88) = -9.94, p < .001, d = -1.57, 95% CI [-2.03, -1.1]
	SE real-exp.	t(179.58) = -0.13, p = .898, d = -0.02, 95% CI [-0.31, 0.27]	t(171.17) = -0.13, p = .898, d = -0.02, 95% CI [-0.32, 0.28]
	SE pred-exp.	t(178.43) = -7.94, p < .001, d = -1.17, 95% CI [-1.49, -0.86]	t(167.89) = -7.93, p < .001, d = -1.2, 95% CI [-1.53, -0.88]
10 pos. vs. 25 pos. words	JE pred-exp.	t(92) = 0.19, p = .847, d = 0.02, 95% CI [-0.18, 0.22]	t(88) = 0.06, p = .948, d = 0.01, 95% CI [-0.2, 0.21]
	SE real-exp.	t(169.45) = 0.24, p = .809, d = 0.04, 95% CI [-0.26, 0.33]	t(163.94) = 0.08, p = .937, d = 0.01, 95% CI [-0.29, 0.31]
	SE pred-exp.	t(180.09) = 1.42, p = .157, d = 0.21, 95% CI [-0.08, 0.5]	t(175.31) = 1.39, p = .167, d = 0.21, 95% CI [-0.09, 0.5]

Note. pred-exp = predicted-experience. real-exp = real-experience.

Exploring potential order effects

All participants took part in both Study 1 and Study 2 in a random order. We tested whether participating in one study influenced the results of the consecutive study. The main inferential analyses reported in the manuscript were based on multiple *t*-tests, one for each evaluation mode (i.e., JE vs. SE) and per comparison of different scenarios (e.g., selling 0 books vs. 80 books, or selling 80 books vs. 160 books etc.). To test whether the order had an effect on happiness ratings, instead of further separate *t*-tests, we ran linear regression models for each study and evaluation mode. These models included dummy variables for the different scenarios (e.g., number of books sold), a dummy variable of participation order (i.e., was the respective study the first one), and the interaction between the scenario dummies and the order dummy. Additionally, we first show a model that ignores the study display order (labeled as Model 1) and only then introduce the order dummy variable and the interactions (labeled as Model 2).

This way, the scenario dummies in Model 2 represent the effect of the different scenarios on happiness ratings when the respective study was the first to be seen, and the interaction terms represent by how much the effect estimates change when the respective study was the second to be seen. A significant interaction term would then signal differences in happiness ratings as a function of display order. If the interaction terms are not statistically significant, we interpret this as absence of any systematic order effects.

The linear models for the SE evaluation mode represent conventional multiple regressions. The models for the JE evaluation mode represent linear mixed effects models with a random intercept for individuals. This accounts for the fact that participants made ratings for each of the

scenarios. Note that all results remain qualitatively the same even when the dependencies are ignored.

Study 1

Table S7 shows results of a linear mixed effects model for happiness ratings in the JE condition of Study 1. The scenario dummies are compared against the "80 copies sold" scenario as reference category. Results of Model 1 express the established pattern of the *t*-test results reported in the main manuscript where happiness in the "80 copies sold" scenario was higher than "0 copies sold" and lower than both "160 copies sold" and "240 copies sold". Adding the study order dummy and the interactions between this dummy with each of the scenario dummies in Model 2 shows that the presentation did not have a statistically significant effect on the happiness ratings.

Table S8 shows results of a linear regression model for happiness ratings in the SE condition of Study 1. Again, results of Model 1 express the *t*-test results where there is only a difference between the qualitatively different scenarios (i.e., selling 0 books vs. 80 books). The interaction terms in Model 2 were not statistically significant, expressing that the presentation order did not have an effect on the happiness ratings.

The happiness ratings for each scenario and condition separated by participation order are plotted in Figure S1, visually confirming the clear similarity between the two different orders.

Table S7
Exploration of order effects for Study 1 JE condition

	Happiness					
	Model 1			Model 2		
Variables	B	SE	p	B	SE	p
Intercept	6.37	0.15	< .001	6.34	0.19	< .001
0 copies sold	-4.36	0.18	< .001	-4.34	0.23	< .001
160 copies sold	1.27	0.18	< .001	1.16	0.23	< .001
240 copies sold	2.03	0.18	< .001	2.03	0.23	< .001
Order				0.06	0.30	.839
0 copies sold*Order				-0.04	0.36	.919
160 copies sold*Order				0.26	0.36	.463
240 copies sold*Order				-0.01	0.36	.980

Note. N = 103. $N_{decisions}$ = 412. Groups were dummy coded with "80 copies sold" as reference category. Order was dummy coded with 0 = "Study 1 first" and 1 = "Study 2 first".

Table S8
Exploration of order effects for Study 1 SE condition

	Happiness					
	Model 1			Model 2		
Variables	B	SE	p	B	SE	p
Intercept	8.16	0.13	< .001	8.02	0.18	< .001
0 copies sold	-5.51	0.18	< .001	-5.35	0.25	< .001
160 copies sold	-0.10	0.18	.564	-0.06	0.25	.816
240 copies sold	-0.14	0.18	.447	0.06	0.25	.805
Order				0.30	0.25	.242
0 copies sold*Order				-0.33	0.36	.355
160 copies sold*Order				-0.09	0.36	.792
240 copies sold*Order				-0.39	0.36	.268

Note. N = 412. Groups were dummy coded with "80 copies sold" as reference category. Order was dummy coded with 0 = "Study 1 first" and 1 = "Study 2 first".

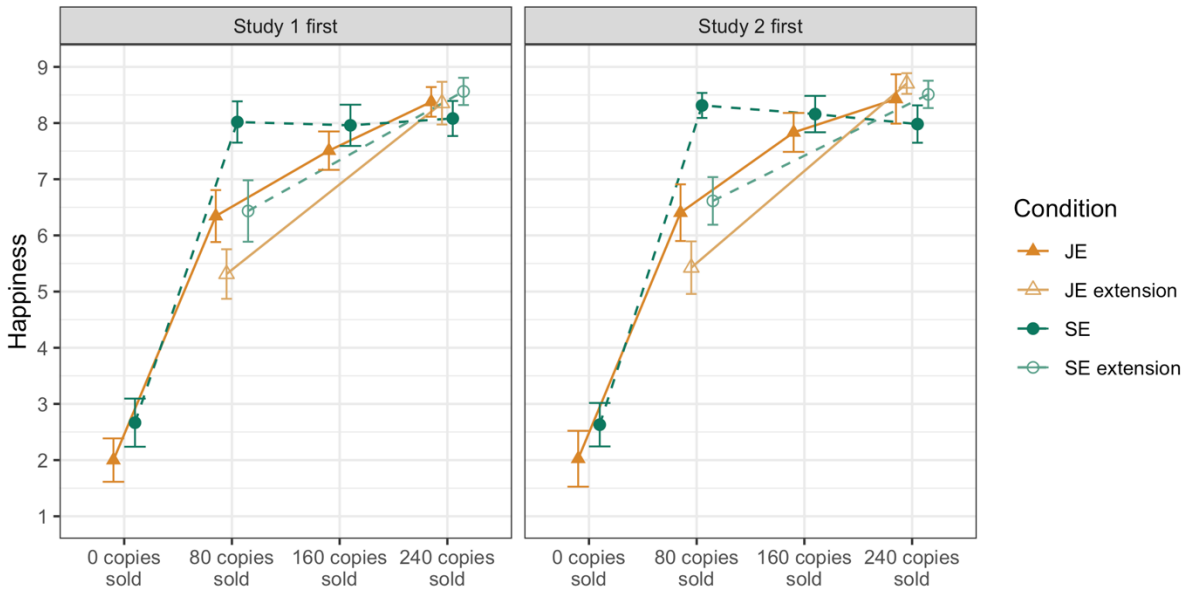

Figure S1. Study 1 happiness ratings by scenario, evaluation mode, and participation order.

Study 2

Table S9 shows results of a linear mixed effects model for happiness ratings in the JE predicted condition of Study 2. The scenario dummies are compared against the "reading 10 negative words" scenario as reference category. Note that the results do not change qualitatively if "reading 10 positive words" is used as reference category. Results of Model 1 express that reading 10 negative words was associated with higher happiness ratings than reading 25 negative words, and lower happiness ratings than in the two positive word scenarios. Importantly, the interaction terms in Model 2 were not statistically significant, suggesting that the presentation order did not affect the results.

Table S10 shows results of a linear regression model for happiness ratings in the SE predicted condition of Study 2. The results follow the same pattern as the JE predicted analysis. As before, none of the interaction terms in Model 2 was significant.

Table S11 shows results of a linear regression model for happiness ratings in the SE real condition of Study 2. As already suggested by the reported *t*-tests, none of the scenarios differed from the reference category. The interaction terms between scenario and presentation order were again not significant.

The happiness ratings for each scenario and condition separated by participation order are plotted in Figure S2, also visually confirming the clear similarity between the two different orders for Study 2.

Table S9
Exploration of order effects for Study 2 JE predicted condition

	Happiness					
	Model 1			Model 2		
Variables	B	SE	p	B	SE	p
Intercept	4.25	0.16	< .001	3.98	0.23	< .001
25 neg. words	-0.69	0.21	.001	-0.49	0.30	.105
10 pos. words	2.23	0.21	< .001	2.51	0.30	< .001
25 pos. words	2.19	0.21	< .001	2.53	0.30	< .001
Order				0.54	0.33	.098
25 neg. words*Order				-0.40	0.43	.348
10 pos. words*Order				-0.58	0.43	.179
25 pos. words*Order				-0.68	0.43	.111

Note. $N = 93$. $N_{decisions} = 372$. Groups were dummy coded with "10 neg. words" as reference category. Order was dummy coded with 0 = "Study 2 first" and 1 = "Study 1 first".

Table S10
Exploration of order effects for Study 2 SE predicted condition

	Happiness					
	Model 1			Model 2		
Variables	B	SE	p	B	SE	p
Intercept	4.75	0.18	< .001	5.00	0.23	< .001
25 neg. words	-0.77	0.25	.002	-0.56	0.34	.105
10 pos. words	1.91	0.25	< .001	2.12	0.34	< .001
25 pos. words	1.57	0.25	< .001	1.50	0.32	< .001
Order				-0.58	0.35	.101
25 neg. words*Order				-0.27	0.49	.581
10 pos. words*Order				-0.25	0.49	.609
25 pos. words*Order				0.15	0.49	.759

Note. $N = 372$. Groups were dummy coded with "10 neg. words" as reference category. Order was dummy coded with 0 = "Study 2 first" and 1 = "Study 1 first".

Table S11
Exploration of order effects for Study 2 SE real condition

	Happiness					
	Model 1			Model 2		
Variables	B	SE	p	B	SE	p
Intercept	5.19	0.18	< .001	5.19	.26	< .001
25 neg. words	-0.08	0.26	.772	0.12	.37	.752
10 pos. words	0.03	0.26	.906	0.17	.37	.658
25 pos. words	-0.04	0.26	.893	0.19	.39	.629
Order				-0.01	.37	.979
25 neg. words*Order				-0.36	.52	.490
10 pos. words*Order				-0.25	.52	.635
25 pos. words*Order				-0.39	.53	.465

Note. $N = 359$. Groups were dummy coded with "10 neg. words" as reference category. Order was dummy coded with 0 = "Study 2 first" and 1 = "Study 1 first".

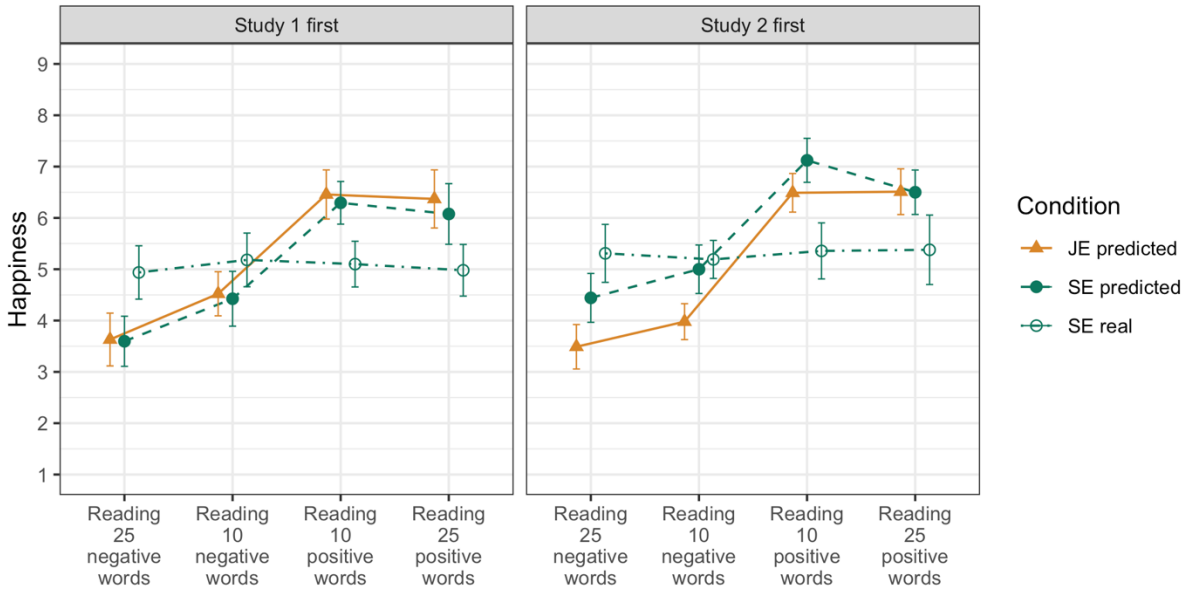

Figure S2. Study 2 happiness ratings by scenario, evaluation mode, and participation order.

References

Hsee, C. K., & Zhang, J. (2004). Distinction bias: Misprediction and mischoice due to joint evaluation. *Journal of Personality and Social Psychology*, *86*(5), 680. https://doi.org/10/dhnrdb

LeBel, E. P., Vanpaemel, W., Cheung, I., & Campbell, L. (2019). A brief guide to evaluate replications. *Meta-Psychology*, *3*. https://doi.org/10/ggfpkc