UNIVERSITY of CALIFORNIA PRESS | Collabra: Psychology

Ziano, I., Yeung, S. K., Lee, C. S., Shi, J., & Feldman, G. (2023). "The Effort Heuristic" Revisited: Mixed Results for Replications of Kruger et al. (2004)'s Experiments 1 and 2. *Collabra: Psychology*, *9*(1). https://doi.org/10.1525/collabra.87489

Social Psychology

"The Effort Heuristic" Revisited: Mixed Results for Replications of Kruger et al. (2004)'s Experiments 1 and 2

Ignazio Ziano[1a], Siu Kit Yeung[2b], Cheong Shing Lee[3c], Jiaxin Shi[4], Gilad Feldman[3d]

[1] Institute of Management, Geneva School of Economics and Management, University of Geneva, Switzerland, [2] Department of Psychology, Chinese University of Hong Kong, Hong Kong SAR, [3] Department of Psychology, University of Hong Kong, Hong Kong SAR, [4] School of Psychology, South China Normal University, Guangzhou, China

Keywords: effort heuristic, judgment and decision making, replication, effort, value

https://doi.org/10.1525/collabra.87489

Collabra: Psychology

Vol. 9, Issue 1, 2023

Kruger, Wirtz, van Boven, and Altermatt (2004) described the effort heuristic as the tendency to evaluate the quality and the monetary value of an object as higher if the production of that object was perceived as involving more effort. We attempted two preregistered replications (total $N = 1405$; U.S. American participants from MTurk and Prolific) of their Experiments 1 and 2. Our first replication using an MTurk sample found support for the original's findings regarding Experiment 2, yet failed to find support for the original's findings in Experiment 1. Our second revised attempt of Experiment 1 on Prolific was mixed, with more nuanced findings, showing support for an effort heuristic effect for liking/quality and no support for an effort heuristic on monetary value. We discuss possible reasons for this discrepancy, theoretical implications and future research directions for the psychology of value and the effort heuristic. All materials, data, and code were made available on https://osf.io/qxf5c/.

Kruger, Wirtz, van Boven, and Altermatt (2004) found that effort information affects judgments of quality. For instance, telling participants that a poem or a painting had taken more time to be finished (e.g., 26 hours vs. 4 hours) increased quality judgments and monetary evaluations, a phenomenon termed "effort heuristic".

We attempted replications of Kruger et al (2004) Experiments 1 and 2[1] in which effort was manipulated between-subjects and within-subjects respectively. We chose this target article due to its academic impact and the lack of close replications of this finding.

We report two replications with mixed results. In our first replication using an Amazon Mechanical Turk (MTurk) sample we found support for Experiment 2 (in our Study 1b), yet failed to find support for Experiment 1 (in our Study 1a). In our second attempt to replicate Experiment 1 using Prolific (our Study 2) we found support for the effort heuristic on liking and quality ratings, but not for monetary value.

Our chosen target: "The Effort Heuristic"

One stream of research emphasized the positive effects of the personal experience of effort. For instance, Olivola and Shafir (2011) showed that personally exerting more effort makes people contribute more to charitable causes, and to public pools in economic games. Norton, Mochon, and Ariely (2012) showed that people tend to like the same products more when they built them themselves, a phenomenon termed "the IKEA effect". These *personal* effort findings are consistent with the theory of cognitive dissonance (Festinger, 1957), meaning that people may try to justify their *personal* efforts and hard work by rationalizing that the quality of products are high and the outcomes are worthwhile (Kruger et al., 2004). How about judgments of *others*' work? Perceptions of effort are also useful in making waiting times more pleasant for customers. For instance, people who are shown a progress bar (an indicator of computational effort) when waiting for a webpage to load report higher satisfaction with the service (Buell & Norton, 2011).

a Contributed equally, joint first author.

b Contributed equally, joint first author.

c Contributed equally, joint first author.

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

1 We refer to "Experiment 1 and 2" when we talk about the original experiments, and we refer to our experiments as "Studies" to differentiate between the original and the replication project.

However, effort does not always increase the perceived quality and value of objects made by third parties. Cho and Schwarz (2008) showed that lay theory of quality can moderate the effort heuristic: if people are convinced that great artistic work requires talent (vs. effort) they are less likely to believe that higher-effort paintings are of higher value. Further, the positive effects of effort are only valuable when human agents exert effort – people prefer slower advice from human agents, because it is considered more effortful and therefore more precise, but the opposite happens for algorithmic recommendation systems (Efendić et al., 2020).

The chosen article for replication on "The Effort Heuristic" by Kruger et al (2004) has been impactful and inspired a flurry of research in psychology in several areas including cognitive, social, economic, and consumer psychology. Overall, the effort heuristic has broader significance for the psychology of evaluation, because it reveals that details of the production process can influence people's evaluations of objects above and beyond evaluations of the features of the object itself. This suggests that evaluations are inherently contextual and depend on how an object is presented, and that effort is one of the factors that people may include in their evaluation. The effort heuristic is also of considerable interest from an economic and consumer psychology standpoint, as it suggests that people have lay theories about value and production, that is, intuitive beliefs of the origin of value in the production process: the higher the effort, the higher the value.

Why would it be worthwhile to replicate Kruger et al. (2004)? First, despite the impact and potential implications in judgment, decision making and consumer psychology, we are not aware of any direct replications of "The Effort Heuristic", and therefore decided to attempt a replication. Kruger et al. (2004) includes three total experiments. We chose to replicate Experiments 1 and 2 because they are the most straightforward demonstrations of the effort heuristic using a between-subjects and within-subjects design respectively (separate evaluation and joint evaluation), whereas in Experiment 3 a moderation (stimuli ambiguity) was introduced. Moreover, previous studies have demonstrated that higher effort or time-related information does not always result in better evaluations (Cho & Schwarz, 2008; Efendić et al., 2020).

Effort Heuristic findings by Kruger et al. (2004)

In Kruger et al. (2004)'s Experiment 1, participants (144 U.S. American undergraduate students) read a poem, called "Order". Participants were randomly assigned to one of two conditions, and they rated the same poem as better when they believed it took 18 hours (d = 0.34 on a liking/quality index) and assigned it a higher monetary value (d = 0.33) compared to when they were told it took 4 hours.

In their Experiment 2, participants (66 U.S. Americans, primarily psychology and art university students) were shown two paintings, named "12 lines" and "Big Abstract". Participants were randomly assigned to one of two conditions, one in which participants were told that "12 Lines" took 26 hours to finish whereas "Big Abstract" took 4 hours to finish and one in which they were told the opposite. In this study, the impact of effort information influenced a liking/quality index and judgments of monetary value. The same painting was better liked and considered to be of higher monetary value when it was indicated as the one that took more effort. This pattern of results yielded interaction effect sizes of η^2_p = 0.09 on a liking/quality index and η^2_p = 0.15 on monetary evaluation. Participants also directly compared the two paintings, both in terms of comparative preference, quality, how an average American would rate their quality, and monetary value. Experiment 2 also included additional features of examining but failing to find support for expertise as a moderating factor, and testing a mediation model of the effort manipulation impacting comparative quality with comparative effort as the mediator.

We provided further details about the results of the target's results in Table 4 of the Supplementary Materials.

Extension: Artistic talent

In addition to replicating the effort heuristic, we were also interested in extending it by learning more about the type of effort-based inferences that people make. Can perception of effort affect perceptions of the artists' talent, beyond judgments of quality of the art object?

It is possible that perceptions of artistic effort and talent have an inverse relationship, such that the higher the perceived effort is, the lower the perceived artistic talent. In fact, Cho and Schwarz (2008) found that a large majority of participants in a pretest agreed with statements such as "Talented producers need to invest less time than untalented producers to achieve the same quality output," and "Unskilled producers need to invest more time and effort in the creation process than skilled producers for the same quality output." This motivates the hypothesis that perhaps a higher amount of invested effort can be a signal of lower innate talent on the part of the producer. On the other hand, it is also possible that perceptions of effort and talent go hand-in-hand, such that artists who expend more effort are perceived as being also more talented. The research about the positive effects of deliberate practice on performance (Ericsson et al., 1993) is now well-known in pop culture (e.g., Gladwell, 2008) and perhaps most people believe that effort is a necessary condition for artistic talent to emerge. This extension may potentially deepen our understanding of the interplay between lay theories of effort and talent, a question initially raised in response to the publication of "The Effort Heuristic" (Cho & Schwarz, 2008), with implications for the evaluation of objects and for the evaluation of labor and workers. We therefore decided to include one additional measure to the replication of Experiment 1 in order to test the two competing hypotheses that we outlined above.

Pre-registrations and open data/code

We first pre-registered the experiment on the Open Science Framework (OSF) and data collection was launched after registration. All materials, dataset, and code were made available on the OSF at https://osf.io/qxf5c/. We report pre-

registration deviations and reasons for deviations in Supplementary Table 20 (p. 52 to p. 54).

All measures, manipulations, and exclusions for this investigation are reported, and data collection was completed before analyses. Pre-registrations were made public on the OSF: https://osf.io/cga6h/ (Study 1) https://osf.io/h6kb2 (Study 2).

Study 1a-1b: Overview of replications with MTurk sample

Participants and sensitivity power analysis

We initially recruited 705 MTurk participants, and according to our preregistered criteria, we excluded a total of 103 participants.[2] This left 602 participants (M_{age} = 39.04, SD_{age} = 12.32, 277 males, 321 females, 4 "Other/would rather not disclose"; all U.S. residents). A sensitivity power analysis using the R package *pwr*, version 1.3 (Champely et al., 2018) showed that we had 98% statistical power detect both of the smallest effect sizes found in the original article (d = 0.33 and V = 0.25), with a two-tailed alpha of 5%.

Procedure: Combined data-collection

All participants read an informed consent form, and then completed replications of Experiment 1 (which we deemed Study 1a) and Experiment 2 (which we deemed Study 1b) in random order.

Study 1a: MTurk Replication of the Original Experiment 1

Method

Stimuli

We used a poem different than the one used in the original, composed by American poet Michael Van Walleghen, "Where She Lives" (1972), which was available under Creative Commons license. We chose this poem as it was written by the same author as the poem in the original article, and in the same time period (the 1970s). The stimuli for all our experiments are available in the Supplementary Materials (pp. 22 to 31).

Procedure

Participants first read the poem. Participants were provided with information regarding the poem, including the title, author, and the age of the author. In line with the original study, participants were then randomly assigned to one of two conditions (high effort and low effort) and were told that the poem had taken either 18 hours or 4 hours to be composed. We then asked four comprehension questions on the information we provided them (in the same page where participants read the poem; available in the Supplementary Materials, p. 22). Participants could not proceed to the rest of the survey until they replied correctly to all comprehension checks. Table 1 includes a classification of this replication following LeBel et al (2018).

Measures

Following the manipulation, participants answered several evaluation questions on the poem in the following order: a) liking of the poem (11-point scale, 1 = *hate it*, 6 = *it's OK* and 11 = *love it*), b) quality of the poem, c) predicted others' quality rating (both on 11-point scales, 1 = *Terrible*, 6 = *OK*, 11 = *Excellent*), d) perceived monetary value (in USD), in which participants enter the amount in an open text box, e) perceived artistic talent of the poet (11-point scale, 1 = *Not talented at all*, 6 = *Mediocre*, 11 = *Extremely talented*). Perceived artistic talent of the poet was absent in the original study, and it was added as part of our extension.

Results

Liking/Quality

We found support for an association between liking and perceived quality, r = .85, p < .001. We averaged liking and quality into a liking/quality index, as in the original paper. In a one-way ANOVA we found no support for the impact of effort information on the liking/quality index (low-effort condition: n = 300, M = 5.26, SD = 2.38; high-effort condition: n = 302, M = 5.14, SD = 2.35, F(1, 599.79) = 0.39, p = .531, η^2_p = 0.00, 90% CI [0.00, 0.01], d = -0.05, 95% CI [-0.21, 0.11]). We provide additional analyses regarding each of the components of the liking/quality index in the Supplementary Statistical Information section in the Supplementary Materials (p. 37).

Monetary value

A Shapiro-Wilk test of normality found the assumption of normality was violated, W = 0.05, p < .001. Therefore, we computed natural-log-transformed values in accordance with the original study and our preregistration. We added 0.5 to all values before transforming.[3] A one-way ANOVA test found no support for differences on the natural-log-transformed perceived monetary value between the two conditions (high-effort condition: M = 3.92, SD = 1.90, $Median_{before\text{-}transformation}$ = \$50; low-effort condition: M = 3.88, SD = 2.10, $Median_{before\text{-}transformation}$ = \$50, F(1,

2 Four participants were excluded due to self-reporting low seriousness (< 4 on 5-Point scale), 33 participants reported they had previously seen or might have seen the artwork, 9 participants reported a low level of English proficiency (self-report level < 5, 7-Point scale), and 57 participants correctly guessed the hypothesis or purpose of the study. Full-sample analyses are reported in the Supplementary Materials (p. 33) and are overall very similar to those after exclusions.

3 We did not pre-register adding 0.5 to all values during pre-registration as we did not anticipate that there would be zeros.

Table 1. Classification of the two replications of the original Experiment 1 (Study 1a) and Experiment 2 (Study 1b) based on LeBel et al. (2018)

Design facet	Study 1a (MTurk)	Details of deviation	Study 1b (MTurk)	Details of deviation
Effect/hypothesis	Same		Same	
Independent Variable construct	Same		Similar	We removed the expertise vs novice independent variable since the original study failed to find support for an expertise effect.
Dependent Variable construct	Same		Same	
Independent Variable operationalization	Same		Same	
Dependent Variable operationalization	Same		Same	
Independent Variable stimuli	Similar	We used a different poem from the same author	Similar	We replaced the paintings with Creative Commons poem/paintings.
Dependent Variable stimuli	Same		Same	
Shared attributed for both Studies 1a and 1b (unified single data collection, random order)				
Procedural details	Similar	We combined Studies 1a and 1b into a unified design in a single data collection, with the studies displayed in random order. The experiment displayed first, mirrors the original design.		
Physical settings	Different	We used an online setting whereas the original used a lab setting		
Population (e.g., age)	Different	We recruited American MTurkers whereas the original study recruited U.S. American college students		
Contextual variables	Different	The replication first replication was conducted in 2018 and the second replication was conducted in 2022 whereas the original was conducted in 2004 or earlier.		
Replication classification	Close Replication		Close Replication	

593.29) = 0.07, p = .786, η^2_p = 0.00, 90% CI [0.00, 0.00], t(600) = 0.27, d = 0.02, 95% CI [-0.14, 0.18]).[4]

Extension – Perceived artistic talent

In a one-way ANOVA we found no support for an effect of effort information on perceived artistic talent (low-effort condition, M = 6.32, SD = 2.33; high-effort condition, M = 6.21, SD = 2.23), F(1, 598.60) = 0.38, p = .54, η^2_p = 0.00, 90% CI [0.00, 0.01], d = -0.05, 95% CI [-0.21, 0.11].

Display order effects

We conducted analyses examining the impact of display order.

Liking/quality

For Experiment 1 examining liking and quality, we found no support for an interaction between effort information and display order, F(1, 598) = 3.36, p = .067, η^2_p = 0.01 [0.00, 0.02]. When Experiment 1 was displayed first, we found no support for differences (higher effort condition: n = 163, M = 5.37, SD = 2.24; lower effort condition: n = 141, M = 5.15, SD = 2.07; t(302) = 0.88, p = .378, d = 0.10 [-0.12, 0.33]). When Experiment 1 was displayed second, we found no support for differences (higher effort condition: n = 139, M = 4.87, SD = 2.46; lower-effort condition: n = 159, M = 5.36, SD = 2.63; t(296) = -1.65, p = .101 , d = -0.19 [-0.42, 0.04]). Though the direction of the effect seems to have reversed, all effects were below the alpha threshold.

Monetary value

For Experiment 1 monetary value, we found no support for an interaction between effort information and display order, F(1, 598) = 3.47, p = .063, η^2_p = 0.01 [0.00, 0.02]. When Experiment 1 was displayed first, we found no support for difference between participants in the higher effort condition (n = 163, M = 4.13, SD = 1.88) and participants in the lower effort condition (n = 141, M = 3.80, SD = 1.92), t(302)

4 For robustness checks, we also conducted alternative calculations, replacing 0 with 0.001 and then log-10-transforming the values, and the results were similar to those above. We made those available in the Supplementary Materials (p. 51).

= 1.53, p = .127, d = 0.18 [-0.05, 0.40]. When Experiment 1 was displayed second, we failed to find support for differences between participants in the higher effort condition (n = 139, M = 3.67, SD = 1.89) and participants in the lower-effort condition (n = 159, M = 3.95, SD = 2.25), t(296) = -1.13, p = .261, d = -0.13 [-0.36, 0.10].

Discussion

Replication

We conclude that this replication of Experiment 1 was unsuccessful. We failed to find support for both liking/quality and monetary value, with effect sizes very close to zero, and confidence intervals not overlapping with the original effect size estimate. The order in which Experiment 1 and Experiment 2 were presented does not seem to explain our results.

Extension: Artistic talent

We failed to find support for the impact of effort information on perceived artistic talent. This may suggest that time spent on this kind of artistic endeavor may not translate to lower perceptions of artistic talent. It is possible that perceptions of effort and talent are not as opposed as Cho and Schwartz (2012) suggested, that our design did not have the correct measure of artistic talent, or that the effect is weaker than we expected, requiring higher statistical power to detect.

Study 1b: MTurk Replication of the Original Experiment 2

Method

Stimuli

We were unsuccessful in obtaining access to the paintings used by the original authors in Experiment 2 ("Big Abstract" and "12 Lines"). We therefore used "The Flirt" (depicting a bird, in color) and "Bold" (depicting a tiger, in black and white), paintings by American artist Barbara Keith in 2014, which we retrieved from https://fineartamerica.com/. We chose these paintings as they were from the same artist, similar to one another in style, not well-known, and were in the common domain. They are provided in the Supplementary Materials (pp. 27 and 29).

Procedure

This study used a 2 (painting: "The Flirt" and "Bold", within-subject) by 2 (paintings with higher effort, "The Flirt" with higher effort vs "Bold" with higher effort between-subject) mixed design. In one condition, participants were told that "The Flirt" had taken 26 hours to finish and "Bold" had taken 4 hours to finish. In the other condition, the effort information was reversed, such that participants were informed that "The Flirt" had taken 4 hours to finish and "Bold" had taken 26 hours to finish.

First, participants were shown one of the two paintings in random order. Participants were told about the title, nationality, equipment, and time taken in completing the painting, in addition to the amount of effort that was our independent variable. Participants answered four comprehension check questions regarding the task (in the same page as the painting - available in the Supplementary Materials, p. 28), and were not allowed to proceed with the experiment until those were answered correctly (with multiple possible attempts).

In the next pages, participants answered four separate evaluation questions on a) liking of the paintings (11-point scale, 1 = *hate it*, 6 = *it's OK* and 11 = *love it*), b) quality of the paintings, c) predicted others' quality rating (for b and c, using 11-point scales, 1 = *Terrible*, 6 = *OK*, 11 = *Excellent*), d) perceived monetary value in USD, for both paintings (in an open text box).

Participants then answered four dichotomous comparative judgments about the paintings: which painting they liked more, which painting had higher quality, which painting would fetch more money in an auction, and which painting had higher average ratings by Americans. Following that, participants were asked comparative rating questions about the paintings: participants compared the paintings with continuous scales, in overall quality and effort invested in (1 = *"The Flirt" much higher* / more, 6 = *exactly the same*, 11 = *"Bold" much higher* / *more*). Finally, they were asked whether they saw the paintings before. A classification of this replication according to LeBel et al (2018) is available in Table 1.

Results

Liking and Quality

Liking, perceived overall quality, and predicted others' quality rating showed high correlation (Cronbach's α = 0.88), and we therefore averaged them in a liking/quality index, as in the original study. We report additional correlations among these measures in the Supplementary Materials (p. 41).

We conducted a 2 (painting: "The Flirt" vs. "Bold", within-subjects) by 2 (higher-effort painting: "The Flirt" vs. "Bold", between-subjects) mixed ANOVA test.[5] "Bold" (M = 8.23, SD = 1.71) was rated higher than "The Flirt" (M = 7.96, SD = 1.66), F(1, 600) = 16.99, p < .001, d = 0.17, 95% CI [0.09, 0.25]. We found support for the predicted two-way interaction between the paintings ratings and paintings with higher effort, F(1, 600) = 10.61, p = .001, η^2_p = 0.02, 90% CI [0.00, 0.04]. When participants were told that "Bold" took a longer time than "The Flirt", "Bold" (M = 8.23, SD = 1.72) was evaluated more positively than "The Flirt" (M = 7.74, SD

5 This analysis represents a deviation from the pre-registered approach, which planned to analyse only effort information (26 hours vs 4 hours) as a factor, yet is more closely aligned with the target article's analyses.

= 1.67), $t(299) = 5.21$, $p < .001$, $d = 0.28$ [0.07, 0.49]. When participants were told that "The Flirt" took a longer time, the difference in evaluation was much smaller ("The Flirt": $M = 8.18$, $SD = 1.63$; "Bold": $M = 8.24$, $SD = 1.70$), $t(301) = 0.61$, $p = .541$, $d = 0.04$ [-0.19, 0.26]. The replication effect was weaker than that of the original study, as the confidence interval around the replication effect size does not include the original point estimate of the effect size.

Monetary value

We computed natural-log-transformed monetary values. We added 0.5, a constant, to all values as we cannot log-transform 0s. We conducted a two-way mixed-model ANOVA using painting and condition as within-subjects and between-subjects factors respectively and found that "The Flirt" ($M = 5.57$, $SD = 1.60$, Median of pre-transformed value = $250) was considered having a higher monetary value than "Bold" ($M = 5.39$, $SD = 1.63$, Median of pre-transformed value = $200), $F(1, 600) = 21.06$, $p < .001$, $\eta^2_p = 0.03$, 90% CI [0.01, 0.06], $d = 0.18$, 95% CI [0.10, 0.26].

We also found support for a two-way interaction between paintings rated and paintings with higher effort, $F(1, 600) = 26.85$, $p < .001$, $\eta^2_p = 0.04$, 90% CI [0.02, 0.07]. When participants were told that "The Flirt" took longer time than "Bold", "The Flirt" ($M = 5.76$, $SD = 1.53$) was rated as having more monetary value than "Bold" ($M = 5.38$, $SD = 1.60$), $t(301) = 6.92$, $p < .001$, $d = 0.40$, 95% CI [0.17, 0.63]. When participants were told that "Bold" took longer time than "The Flirt", the difference in value was minimal ("The Flirt": $M = 5.39$, $SD = 1.65$; "Bold": $M = 5.41$, $SD = 1.66$), $t(299) = 0.42$, $p = .676$, $d = 0.02$, 95% CI [-0.20, 0.25]. The effects were weaker than that of the original study.

Dichotomous comparative judgments

Based on the original study, we expected that there would be an effect of effort condition (painting with higher effort) on comparative judgement for all four dependent variables (liking, quality, predicted others' quality rating, and monetary value). We conducted a series of tests for equality of proportions. These results show that the effort heuristic was supported and sizeable for all of the dependent variables, with the exception of liking, for which, however, the effect was in the expected direction. We report them in detail below and summarized in Table 22 in the Supplementary Materials (p. 56).

Liking

When "Bold" was the higher effort painting, 126/300 (42%) participants preferred "The Flirt" over "Bold". When "The Flirt" was the higher effort painting, 143/300 (47%) participants preferred "The Flirt" over "Bold", $\chi^2 = 1.74$, $p = .187$, Cramer's V = 0.05, 95% CI [0.00, 0.13].

Quality

When "Bold" was the higher effort painting, 131/300 (44%) participants indicated that "The Flirt" had higher quality than "Bold". When "The Flirt" was the higher effort painting, 178/300 (59%) participants indicated that "The Flirt" had higher quality than "Bold", $\chi^2 = 14.05$, $p < .001$, Cramer's V = 0.15, 95% CI [0.07, 0.23].

Predicted others' quality rating

When "Bold" was the higher effort painting, 156/300 (52%) participants indicated that others would believe that "The Flirt" had higher quality than "Bold". When "The Flirt" was the higher effort painting, 198/300 (66%) participants indicated that others would believe "The Flirt" had higher quality than "Bold", $\chi^2 = 11.43$, $p < .001$, Cramer's V = 0.14, 95% CI [0.06, 0.22].

Monetary value

When "Bold" was the higher effort painting, 147 /300 (49%) participants indicated that "The Flirt" had higher monetary value than "Bold". When "The Flirt" was the higher effort painting, 199/300 (66%) participants indicated that "The Flirt" had higher monetary value than "Bold", $\chi^2 = 17.57$, $p < .001$, Cramer's V = 0.17, 95% CI [0.09, 0.25].

Continuous comparative judgments

We also found support for the effects of effort information on the two continuous comparative measures of effort (which are akin to manipulation checks) and quality. When "Bold" was the higher effort painting, participants were more likely to indicate that "Bold" took more effort ($M = 8.08$, $SD = 2.65$) compared to when the higher effort painting was "The Flirt" ($M = 4.08$, $SD = 2.83$; $d = 1.46$, $p < .001$, 95% CI [1.26, 1.66]).

When "Bold" was the higher effort painting, participants were more likely to indicate that the higher quality painting was "Bold" (M = 6.46, SD = 2.59) compared to when the higher effort painting was "The Flirt" (M = 5.78, SD = 2.68; $d = 0.26$, $p = .002$, 95% CI [0.10, 0.42]).

Mediation analysis

We used the "medmod" module in JAMOVI (2020) to conduct a mediation analysis using condition as the independent variable, the comparative effort measure as the mediator, and the comparative quality measure as the dependent variable. We found support for an indirect effect of effort information on perceived quality through perceived effort, ab (SE) = -2.15 (0.18), 95% CI [-2.50, -1.80], z = -12.16, $p < .001$, as in the original article, which is in support of the hypothesis that increasing perceived effort increases perceptions of quality.[6]

6 This mediation analysis was not pre-registered yet follows on the analyses conducted in the target paper, which used Sobel test to test the same mediational hypothesis.

Table 2. Study 1b: Comparative judgments

Study	Replication Effect size and CI	Original Effect size and CI	Interpretation (based on LeBel et al., 2019)
Dichotomous Comparative Judgments	Liking: *V* = 0.05, 95% CI [0.00, 0.13] Quality: *V* = 0.15, 95% CI [0.07, 0.23] Others' Quality: *V* = 0.14, 95% CI [0.06, 0.22] Monetary Value: *V* = 0.17, 95% CI [0.09, 0.25]	All 4 Dependent Variables: *V* > 0.25	The original only reported chi squared > 4.1 for all 4 Dependent Variables so magnitude cannot be compared. We found signal for quality, perceived others' rating, and monetary value, but no signal for liking.
Continuous comparative judgment in quality	*d* = 0.26, 95% CI [0.10, 0.42]	*d* = 0.56, 95% CI [0.06, 1.06]	Signal, inconsistent, smaller

Display order effects

Liking/quality index

For Experiment 2 examining liking and quality, we found no support for an interaction between effort information and display order, $F(1, 598) = 0.26$, $p = .609$, $\eta^2_p = 0.00$ [0.00, 0.01]. The effects of Experiment 2 displayed first (Liking and Quality Interaction with Effort Condition: $\eta^2_p = 0.01$ [0.00, 0.04], $p = .056$) were similar to the results of Experiment 2 displayed second (Displayed Second: Liking and Quality Interaction with Effort Condition: $\eta^2_p = 0.02$ [0.00, 0.06], $p = .008$), yet we note that the effects when Experiment 2 was displayed first were close to our alpha threshold (.05).

Monetary value

For Experiment 2 monetary value, we found no support for an interaction between painting rated, effort information (paintings with higher effort), and display order, $F(1, 598) = 0.29$, $p = .589$, $\eta^2_p = 0.00$ [0.00, 0.01]. The results of Experiment 2 displayed first (Monetary Value Interaction with Effort Condition: $\eta^2_p = 0.04$ [0.01, 0.08], $p < .001$) were similar to the results of Experiment 2 displayed second (Monetary Value Interaction with Effort Condition: $\eta^2_p = 0.05$ [0.02, 0.09], $p < .001$).

We found larger effects of effort information in dichotomous comparative judgment when Study 1b was displayed second compared to when Study 1b was displayed first, though we are unsure as to how to interpret that given that we found little to no support for presentation order effects,

Discussion

We successfully replicated most findings of Experiment 2. When participants were told that "The Flirt" took higher effort, "The Flirt" was evaluated similarly to "Bold" on a range of liking and quality measures, with comparative or independent ratings. When participants were told that "Bold" took longer time than "The Flirt", participants evaluated "Bold" in a more positive fashion compared to "The Flirt". Similarly, when "Bold" was the higher-effort painting, there was little difference in monetary evaluation between "Bold" and "The Flirt", but when "The Flirt" was higher-effort, "The Flirt" was considered as having higher monetary value. The comparative judgments analyses yielded similar results consistent with the original, albeit with smaller effect sizes. Of note, an ideal comparison between the original dichotomous judgments and the replication dichotomous judgments is not possible, as there was not enough statistical information provided in the original article (see Table 2). Overall, for the effects that could be compared to the original's, we obtained weaker effects, as the confidence intervals around the effects we found do not include the point estimates of the effect sizes found in the original article.

Study 2: Prolific Replication of the Original Experiment 1

Following the unsuccessful attempt to replicate Experiment 1, we conducted a second attempt with revisions and on a different online platform. We classified this second attempt as a very close replication of the original Experiment 1 (see Table 3).

Methods

Participants

We recruited 700 U.S. American participants from Prolific, of which all passed the attention check "Have you ever

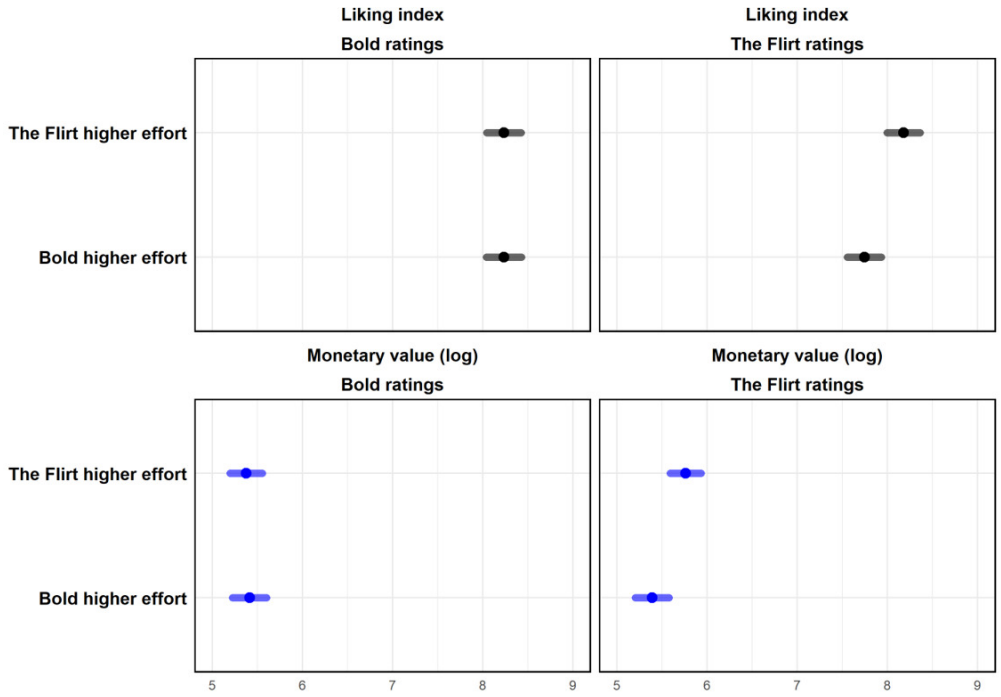

Figure 1. Results of the replication of Experiment 2 on MTurk (Study 1b). Means and 95% confidence intervals are depicted.

been on the planet Venus?" at the start of the survey[7] (210 males, 480 females, 10 other gender; M_{age} = 35.13, *SD* = 12.38). A sensitivity analysis indicated 99.2% power to detect the original effect size *d* = 0.33 for monetary value and 99.8% power to detect the original effect size *d* = 0.34 for liking/quality with an alpha level of 5% (two-tailed).

Procedure

Participants were randomly assigned to one of two conditions, low and high effort. In both conditions, they were given information about a poem which then they read ("Order" by Michael van Walleghem). In the low effort condition, they were told that the poet composed the poem in 4 hours, while in the high effort condition they were told that the poet composed the poem in 18 hours. After, participants completed a series of comprehension checks. They could not advance until they replied correctly. Then, participants completed measures about their liking of the poem and their quality assessment as in previous studies (both anchored at 1 and 11 - these two items were averaged as they showed a high correlation, *r* = 0.77, *p* < .001), and were asked to indicate in a textbox how much money the poem would fetch (in USD) if sold to a poetry magazine. Participants were asked how much time and how much effort they believed that the poem took to be composed, on two items anchored at 1 (*very little time/effort*), 6 (an average amount of time/effort) and 11 (*an extreme amount of time/effort*).[8] These two items were averaged as they showed a high correlation (*r* = 0.64, *p* < .001). The stimulus and all these items were presented in the same page. In the next pages, participants answered funneling and debriefing, in which they were asked if they had seen similar materials before, and finally provided feedback and demographic information.

Results and Discussion

Time/Effort

Participants rated the poem as taking more time and effort in the high effort condition (*n* = 356, *M* = 7.72, *SD* = 2.00) compared to the low effort condition (*n* = 344, *M* = 5.98, *SD* = 1.76), *t*(698) = 12.24, *p* < .001, *d* = 0.93, 95% CI [0.76, 1.09], confirming a successful manipulation.[9]

Liking/Quality

Participants rated the poem as higher quality in the high effort condition (*M* = 6.73, *SD* = 1.95) compared to the low effort condition (*M* = 6.31, *SD* = 1.78; *t*(698) = 3.00, *p* = .003, *d* = 0.23, 95% CI [0.08, 0.38], see Figure 2).

7 We included the attention check at the start of the survey, before randomization, at the suggestion of a reviewer, to avoid differential attrition across conditions (Zhou & Fishbach, 2016)

8 These two items about time and effort were not included in the original study, and were included as a manipulation check.

9 For all measures, results were very similar (regarding descriptive statistics, effect sizes, and p-values – see supplementary materials, p. 34 and p. 35) when excluding participants based on pre-registered exclusion criteria. .

Table 3. Classification of the Prolific replication of the original Experiment 1 (our Study 2), based on LeBel et al. (2018)

Design facet	Study 2 (Prolific)	Details of deviation
Effect/hypothesis	Same	
Independent Variable construct	Same	
Dependent Variable construct	Same	
Independent Variable operationalization	Same	
Dependent Variable operationalization	Same	
Independent Variable stimuli	Same	
Dependent Variable stimuli	Same	
Procedural details	Same	
Population	Different	We recruited U.S. American Prolific participants; the original study recruited U.S. undergraduates
Physical settings	Different	We used an online setting whereas the original used a lab setting
Contextual variables	Different	The replication was conducted in 2021 whereas the original was conducted in 2004 or earlier.
Replication classification	Very close replication	

Monetary value

As preregistered, we log-transformed monetary value. We found no support for difference between log-transformed monetary value in the low effort (M = 5.01, SD = 1.87) and the high effort conditions (M = 4.86, SD = 1.88; $t(698) = 1.00$, $p = .32$, $d = 0.08$, 95% CI [-0.07, 0.22]).[10]

Overall, we found support for the effort heuristic findings for liking and quality, as the 95% confidence interval around the replication effect size included the original point estimate. This finding differs from the results of Study 1a, in which we found no evidence for the effort heuristic. We did not find support for effort heuristic on monetary value, as the 95% confidence interval around the replication effect size did not include the original point estimate. This finding is consistent with our failure to find an effect in Study 1a. We discuss both in detail in the General Discussion.

General Discussion

We summarize our findings and interpretations of the replications in Table 4. We found mixed support for Kruger et al. (2004)'s findings on "the effort heuristic". In our initial study using MTurk we found support for the original's findings for the original Experiment 2 (in our Study 1b) but not for Experiment 1 (in our Study 1a). A second attempt of replicating Experiment 1 using a sample on Prolific (our Study 2) yielded mixed findings, with support for liking/quality and no support for monetary value. The findings of the MTurk and Prolific samples were similar regarding monetary value but diverged on liking/quality, and given that there were several adjustments we are unsure as to which may have led to the different results (see Table 5 for an overview). In our extension to Study 1a we found no support for an effect of effort information on perceived artistic talent. In our extension to Study 2 we found that even in separate decision mode, effort information had a medium-to-large effect on perceptions of effort.

We are unsure as to the reason for the inconsistent findings between the target's Experiment 1, Study 1a replication, and Study 2 replication. In both replication attempts we observed weaker effects compared to the original. Below, we discuss possible reasons for the discrepancy in finding the effort heuristic, offering some theoretical implications and future avenues for the study of the phenomenon.

Samples and participants

The original authors recruited U.S. American students, on-campus, to participate in their study, whereas we recruited online MTurk and Prolific U.S. residents. This was a needed pre-registered deviation, yet we do not believe that the sample is the cause for the mixed findings. First, in our joint replication of Experiments 1 and 2 on MTurk we found support for the original's Experiment 2 but not for the original's Experiment 1, in a combined data collection using the same participants. Second, to address at-

10 Exploratory non-parametric analyses on the non-transformed monetary value showed similar results, Mann-Whitney U = 57353.5, p = .15, and the median values in both conditions were $100.

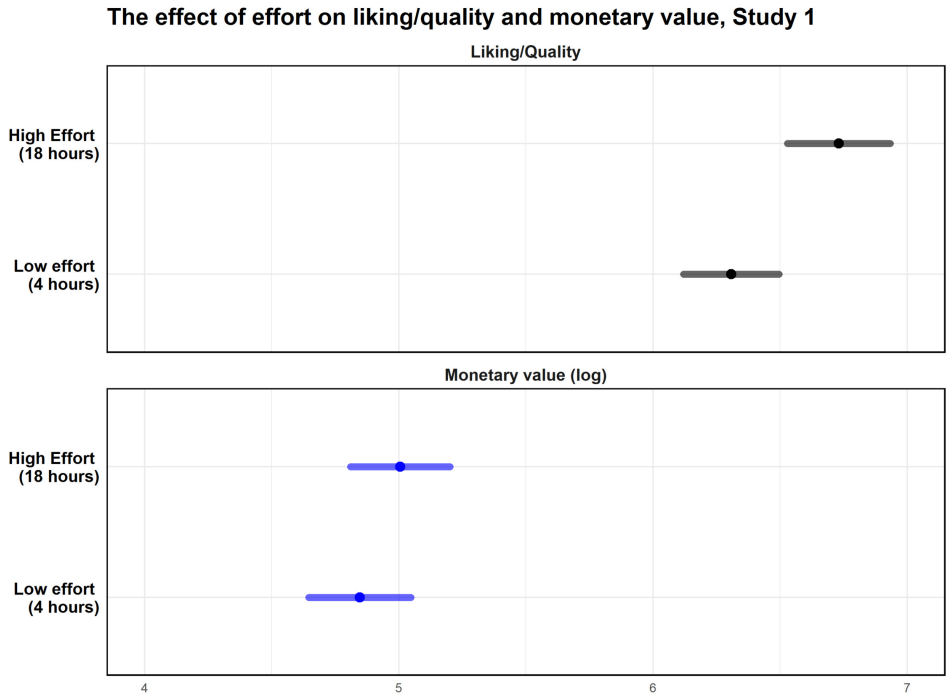

Figure 2. Results of the Prolific replication of Experiment 1 (Study 2). Means and 95% confidence intervals are depicted.

Table 4. Comparison between original and replication effect sizes

Experiment	Dependent variable	Original effect size [95% or 90 % CI]**	MTurk sample replication effect size [95% or 90% CI]**	Interpretation following LeBel et al. (2019)	Prolific sample replication effect size [95% CI]	Interpretation following LeBel et al. (2019)
Experiment 1 replication	Liking/quality index	d = 0.34, [0.00, 0.68]	d = -0.05 [-0.21, 0.11]	No signal, inconsistent	d = 0.23 [0.08, 0.38]	Signal, consistent
	Monetary value	d = 0.33, [-0.02, 0.67]	d = 0.02 [-0.14, 0.18]	No signal, inconsistent	d = 0.08 [-0.07, 0.22]	No signal, inconsistent
Experiment 2 replication	Liking/quality index*	η^2_p = 0.09, [0.01, 0.21]	η^2_p = 0.02, [0.00, 0.04]	Signal, inconsistent, smaller	-	-
	Monetary value*	η^2_p = 0.15, [0.03, 0.28]	η^2_p = 0.04, [0.02, 0.07]	Signal, inconsistent, smaller	-	-

Note. *Interaction effect. **90% CI for η^2_p which cannot be negative. More detailed information is contained in Table 23 in the Supplementary Materials.

tentiveness and seriousness we found very similar results comparing the full sample (705 participants) and the sample after exclusion (602 participants; findings are detailed in the Supplementary Materials pp. 32-36). Third, similar studies in judgment and decision-making originally conducted with U.S. American undergraduate students have been replicated successfully with U.S. online workers (see e.g., Jung et al., 2019; Ziano et al., 2021; Ziano, Kong, et al., 2020; Ziano, Mok, et al., 2020).

Stimuli and procedure

In Study 1a (MTurk) we used a different poem compared to the original, and a procedure in which participants first saw the poem, and then made judgments in subsequent pages. In our second attempted replication of Experiment 1 using Prolific (Study 2) we adjusted to use the same exact poem as in the original study, and to display the poem and the dependent variables together in the same page. Findings were closer to that of the original yet still mixed; we found support for liking/quality, but not for monetary value. In Study 1b, we used different stimuli from the original Experiment 2, by choosing naturalistic depictions of animals rather than abstract paintings. This may be the reason of the weaker effects we found in the Experiment 2 replication. As Kruger et al (2004) demonstrated in Experiment 3, the effort heuristic was stronger when the object was harder to evaluate (i.e. when the stimulus became more

Table 5. Differences between the MTurk replication of Experiment 1 (Study 1a) and the Prolific replication of Experiment 1 (Study 2)

Design facet	MTurk replication of Experiment 1 (Study 1a)	Prolific replication of Experiment 1 (Study 2)
Sample size	705 U.S. American MTurk participants (602 after exclusions)	700 U.S. American Prolific participants
Independent Variable stimuli	"Where she lives" by Michael van Walleghen	"Order" by Michael van Walleghen
Dependent Variable stimuli	Same as the original (with artistic talent addition)	Same as the original (with perceived time and effort addition)
Procedural details	Participants answered both Study 1a and Study 1b stimuli (in random order)	Participants completed the survey without being exposed to other stimuli first
Page presentation	Stimuli and measures were shown across different pages	Stimuli and measures were shown on the same page
Physical settings	Online	Online
Contextual variables	Replication conducted in 2018	Replication conducted in 2021

ambiguous because it was presented in lower resolution), and it might be that the naturalistic paintings (i.e., paintings of animals) that we used were easier to evaluate than abstract paintings.

There were several differences between Studies 1a-1b and Study 2 replication in terms of sample (MTurk versus Prolific), the combination with replication of the original's Experiment 2, the stimuli used, and display format (same versus different page). Though findings regarding monetary value were consistent (and consist of an unsuccessful replication), any of these factors may explain the divergence in findings regarding quality/liking. Of all these factors, it is possible that the effect is sensitive to the stimuli used, and future investigations are needed to compare the original's to new stimuli to better ascertain the impact of stimuli.

In Study 1a, we used a between-subjects mode, whereas in Study 1b we used a within-subjects mode. However, in Study 2, we found some support for the effort heuristic using a between-subjects mode (separate evaluation). Can evaluation mode be responsible of the detection of the effort heuristic? We do not have the right data to answer this question as, in addition to evaluation mode, there are other differences between the two replication attempts. Future research can directly compare joint and separate evaluation mode to test whether the effort heuristic is stronger in one or the other.

Time

Our replications were conducted in 2018 (Studies 1a and 1b) and 2022 (Study 2), whereas the original authors collected their data on or before 2004. We believe that it is unlikely that the difference in time passed caused the discrepancy between our results and the original results. First, there are many recent papers reporting successful replications of judgment and decision making papers originally published around or before 2004 (e.g., Chen et al., 2020; Ziano, Kong, et al., 2020; Ziano, Wang et al., 2020). Second, it is difficult to reconcile this explanation with the fact that though the replications of Experiment 1 were weak and mixed, the replication of Experiment 2 was successful, with the first data collection in Study 1 conducting the two experiments by the same participants from the same sample. As in the case of sample as a reason for the result discrepancy, we believe that, even if this explanation was tenable, it would be important to know that the effort heuristic may have changed in the last 15 years. Further, we found partial support for effort heuristic for liking/quality of Experiment 1 in our Study 2, which makes time even less likely as the possible explanation.

Order effects

Our experimental procedure (for Studies 1a and 1b only) employed the same sample for completing the replications of both the original Experiment 1 and Experiment 2, presented in random order as our Studies 1a and 1b. However, it seems unlikely that this explains the mixed results, as Experiment 1 was an unsuccessful replication attempt whereas Experiment 2 was successful, regardless of order. Testing for the possibility of order, we did not find support for order effects in either replication, making order effects less likely (though null effects should be interpreted with caution). That we found some support for an effect of the effort heuristic as presented in the original Experiment 1 (between-subjects) in our Study 2, further reduces the likelihood that order effects were crucial in reproducing the effect. However, we cannot completely rule out the possibility of order effects impacting our findings.

Theoretical Implications and Future Research

In evaluating targets in effort heuristic paradigms, participants rely on two essential elements: effort information and the stimulus itself. We discuss the theoretical implications of these two factors for the effort heuristic and suggest two avenues for the systematic investigation of the evaluability of the stimuli and of effort information that may prove fruitful for future research. We focused our discussion on the diverging results of the liking/quality measure (both failed to find support for the monetary value measure). We note that both the original effects (e.g., d = 0.34) and the successfully replicated effects (e.g., d = 0.23)

are considered rater small (Xiao et al., 2023), and therefore very large samples are likely needed to credibly detect interaction effects and moderators of effort heuristic.

Stimuli evaluability

The original authors, in Experiment 3, pointed to stimulus ambiguity as a moderating factor of the effort heuristic, by showing that the effort heuristic was stronger when a picture of an armor suit was presented with a lower image resolution. They argued that when the stimulus was less ambiguous, participants were more likely to rely on the features of the stimulus, but when the stimulus was more ambiguous, participants were more likely to rely on effort information. Their notion of "ambiguity" resembles the notion of "evaluability", that is, the ease with which a value or a comparison of values can be mapped onto an evaluation, or how easy it is for individuals to understand an appraise a stimulus (Hsee & Zhang, 2010; Lembregts & Van Den Bergh, 2019).

The original authors' suggestion can be extended, by examining whether stimulus evaluability may be a factor necessary in order to observe it, rather than a moderator. A possible way to tackle this issue more systematically would be to compare the impact of the effort heuristic on stimuli that differ in evaluability in large-scale studies.

Effort information evaluability

Our successful replication of Experiment 2 suggests that effort information might be hard to evaluate. Unless participants have sufficient experience in assessing similar stimuli, it is unlikely that they would be able to evaluate the effort put in the work and its associated value. In our Study 2 (replication of the original Experiment 1), we added two factors that might have increased ease of evaluation of effort information. First, participants rated the poem in the same page where they read it. In this way, they had access to effort information when they were rating the stimulus, likely making it easier to evaluate. Second, we asked participants to rate how much time and effort they believed the poet took in composing the poem. This might have focused participants' attention on effort information, which in turn might have made it easier to evaluate. Note that if this is correct, and ease of evaluation is necessary in order to observe an effort heuristic, this enriches our understanding of the effort heuristic because it points to a factor that gives way to it. At the same time, this would also reduce somewhat the import of the original findings as it points out that the effort heuristic is not as pervasive as originally formulated. Future research might test this suggestion by adding or subtracting these factors and by measuring ease of effort information evaluation, though this might require very large samples (Giner-Sorolla, 2018).

Strategies that aim at testing this hypothesis could find ways to make effort information easier to evaluate (i.e., less ambiguous, easier to evaluate), for instance by isolating the effect of effort information and making the comparison between joint and separate evaluation. Another way to get at the same objective is to give participants a benchmark that makes effort information easier to understand even in separate evaluation. For instance, a future researcher may try and make effort information easier to understand by providing participants with a reference point that makes the same amount of effort seem larger or smaller in separate evaluation. Another way would be to manipulate evaluation mode, such that a group of participants would be randomly assigned to either evaluate one painting (separate evaluation) or to evaluate several paintings at a time (joint evaluation). In the separate evaluation condition, the experimenter may manipulate effort information between-subjects, whereas in the joint evaluation condition the experimenter would manipulate effort information within-subjects. Joint evaluation mode typically makes information (in this case, the amount of effort) easier to evaluate (Hsee & Zhang, 2010), by offering participants a benchmark with which they can be compared. The issue of different evaluation modes is also potentially important for real-life consumer judgment and decision making, as consumers sometimes compare qualities of two products and/or choose between two products, or simply evaluate a product and decide whether they would like to buy a product occasionally.

Monetary Value

We also observed inconsistent results across different paradigms regarding our second dependent variable of interest, monetary value. For replications of the original Experiment 1 (i.e., our Study 1a and Study 2), we did not find support for effort heuristic on monetary value. However, in the replication of the original Experiment 2 (i.e., our Study 1b), we found support for a smaller effect that was consistent with the original effect regarding the main measure of monetary value, and we found effects that seem consistent with the original regarding the comparative measures. We are unsure of the reason for these differences, and unfortunately, testing them is beyond the scope of the present research. There were many differences between Experiment 1 and 2, including stimuli, effort information, and presentation mode. Each of these factors could by itself be responsible for the difference in results, but we do not have the data to answer this question. We hope that future research could further systematically examine varying these factors to examine moderating effects.

Kruger et al. (2004) and our replications focused on quality/liking and monetary evaluation, but did not include measures with more real-life implications such as intentions and/or decisions to consume. These variables are likely positively correlated, but simply liking the product because creating the product requires higher effort does not necessarily translate into getting the product. It also seems plausible that monetary evaluation would be more strongly correlated with consumer intentions and decisions, compared to the correlations between liking/quality and consumer intentions and decisions. Future studies may measure these additional variables.

Conclusion

The present study attempted three close replications of the effort heuristic and yielded mixed results. We discussed several explanations which seem inconsistent with the data we presented, including differences in stimuli, time, sample, and order effects. We call for future research to systematically investigate factors that may impact the strength of the effort heuristic effect, starting from the evaluability of stimuli, effort information, and additional intention and decision measures. We found some support for effort heuristic, that is, that there is an effect of effort information on liking and quality, and on perceptions of monetary value. However, our replication results also suggest that the effect is possibly weaker and more contextual than previously thought. This may imply that effort heuristic would also have weaker impact on real-world situations, making it more difficult and complex to capture, understand, and address in real-life.

Competing Interests

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

Funding

The research was supported by the European Association for Social Psychology Seedcorn grant awarded to the corresponding author.

Acknowledgments

We thank Tony Evans for his kind and useful comments on an earlier version of this draft.

Contributions

Siu Kit and Cheong Shing conducted the studies. Jiaxin guided and commented on an earlier draft. Gilad led the replication effort, supervised each step in the project, conducted the pre-registrations, and ran data collection. Ignazio and Siu Kit followed up on the initial work by the other coauthors to verify analyses and conclusions as well as to perform new analyses. They also completed the manuscript submission draft. Ignazio, Siu Kit, and Gilad jointly finalized the manuscript for submission.

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

Submitted: September 16, 2022 PDT, Accepted: June 30, 2023 PDT

Role	Ignazio Ziano	Cheong Shing Lee	Siu Kit Yeung	Jiaxin Shi	Gilad Feldman
Conceptualization					X
Pre-registration		X	X		
Data curation					X
Formal analysis	X	X	X		
Funding acquisition					X
Investigation	X	X	X		X
Methodology		X	X		X
Pre-registration peer review / verification	X	X	X	X	X
Data analysis peer review / verification	X		X		
Project administration					X
Resources					X
Software	X	X	X		
Supervision	X			X	X
Validation	X	X	X	X	
Visualization	X		X		
Writing-original draft	X	X	X		
Writing-review and editing	X		X		X

References

Buell, R. W., & Norton, M. I. (2011). The Labor Illusion: How Operational Transparency Increases Perceived Value. *Management Science*, *57*(9), 1564–1579. https://doi.org/10.1287/mnsc.1110.1376

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

Chen, J., Hui, L. S., Yu, T., Feldman, G., Zeng, S., Ching, T. L., Ng, C. H., Wu, K. W., Yuen, C. M., Lau, T. K., Cheng, B. L., & Ng, K. W. (2020). Foregone Opportunities and Choosing Not to Act: Replications of Inaction Inertia Effect. *Social Psychological and Personality Science*, *12*(3), 333–345. https://doi.org/10.1177/1948550619900570

Cho, H., & Schwarz, N. (2008). Of great art and untalented artists: Effort information and the flexible construction of judgmental heuristics. *Journal of Consumer Psychology*, *18*(3), 205–211. https://doi.org/10.1016/j.jcps.2008.04.009

Efendić, E., van de Calseyde, P. P. F. M., & Evans, A. M. (2020). Slow response times undermine trust in algorithmic (but not human) predictions. *Organizational Behavior and Human Decision Processes*, *157*(January), 103–114. https://doi.org/10.1016/j.obhdp.2020.01.008

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406. https://doi.org/10.1037/0033-295x.100.3.363

Festinger, L. (1957). *A theory of cognitive dissonance* (Vol. 2). Stanford university press.

Giner-Sorolla, R. (2018). *Powering Your Interaction*. Approaching Significance. https://approachingblog.wordpress.com/2018/01/24/powering-your-interaction-2/

Gladwell, M. (2008). *Outliers: The story of success*. Little, Brown.

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

Kruger, J., Wirtz, D., Van Boven, L., & Altermatt, T. W. (2004). The effort heuristic. *Journal of Experimental Social Psychology*, *40*(1), 91–98. https://doi.org/10.1016/s0022-1031(03)00065-9

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

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

Lembregts, C., & Van Den Bergh, B. (2019). Making each unit count: The role of discretizing units in quantity expressions. *Journal of Consumer Research*, *45*(5), 1051–1067. https://doi.org/10.1093/jcr/ucy036

Norton, M. I., Mochon, D., & Ariely, D. (2012). The IKEA effect: When labor leads to love. *Journal of Consumer Psychology*, *22*(3), 453–460. https://doi.org/10.1016/j.jcps.2011.08.002

Olivola, C. Y., & Shafir, E. (2011). The Martyrdom Effect: When Pain and Effort Increase Prosocial Contributions. *The Journal of Behavioral Decision Making*, *26*(December 2011), 91–105. https://doi.org/10.1002/bdm

Xiao, Q., Yeung, S. K., Dunleavy, D. J., Röseler, L., Feldman, G., Elsherif, M., & Feldman, G. (2023). *Effect sizes and confidence intervals guide*. https://doi.org/10.17605/OSF.IO/D8C4G

Zhou, H., & Fishbach, A. (2016). The pitfall of experimenting on the web: How unattended selective attrition leads to surprising (yet false) research conclusions. *Journal of Personality and Social Psychology*, *111*(4), 493–504. https://doi.org/10.1037/pspa0000056

Ziano, I., Kong, M. F., Kim, H. J., Liu, C. Y., Wong, S. C., Cheng, B. L., & Feldman, G. (2020). Revisiting Disjunction Effect: Replication of Tversky and Shafir (1992) and extension comparing between and within subject designs. *Journal of Economic Psychology*, *53*(9), 1689–1699. https://doi.org/10.1017/cbo9781107415324.004

Ziano, I., Mok, P. Y. (Cora), & Feldman, G. (2020). Replication and Extension of Alicke (1985) Better-Than-Average Effect for Desirable and Controllable Traits. *Social Psychological and Personality Science*, *June*. https://doi.org/10.1177/1948550620948973

Supplementary Materials

Peer Review History

Download: https://collabra.scholasticahq.com/article/87489-the-effort-heuristic-revisited-mixed-results-for-replications-of-kruger-et-al-2004-s-experiments-1-and-2/attachment/178908.docx?auth_token=6cvO2OKbdQuvyLKS5j8I

Supplementary Materials

Download: https://collabra.scholasticahq.com/article/87489-the-effort-heuristic-revisited-mixed-results-for-replications-of-kruger-et-al-2004-s-experiments-1-and-2/attachment/178909.docx?auth_token=6cvO2OKbdQuvyLKS5j8I