We are all less risky and more skillful than our fellow drivers: Successful replication and extension of Svenson (1981)

Lina Koppel[1], David Andersson[1], Gustav Tinghög[1,2], Daniel Västfjäll[3,4], and Gilad Feldman[5]

[1]Department of Management and Engineering, Division of Economics, Linköping University
[2]Department of Health, Medicine and Caring Sciences, Division of Health Care Analysis, Linköping University
[3]Department of Behavioral Sciences and Learning, Division of Psychology, Linköping University
[4]Decision Research, Eugene, OR
[5]Department of Psychology, University of Hong Kong, Hong Kong SAR

The better-than-average effect refers to the tendency to rate oneself as better than the average person on desirable traits and skills. In a classic study, Svenson (1981) asked participants to rate their driving safety and skill compared to other participants in the experiment. Results showed that the majority of participants rated themselves as far above the median, despite the statistical impossibility of more than 50% of participants being above the median. We report a preregistered, well-powered (total N = 1,203), very close replication and extension of the Svenson (1981) study. Our results indicate that the majority of participants rated their driving skill and safety as above average. We added different response scales as an extension and findings were stable across all three measures. Thus, our findings are consistent with the original findings by Svenson (1981). Materials, data, and code are available at https://osf.io/fxpwb/.

Keywords: better-than-average effect, self-evaluation, self-enhancement, replication

When people are asked to rate themselves on desirable traits and skills, most people rate themselves as above average. This is known as the better-than-average effect and has been demonstrated in a variety of domains. In one of the most well-known examples, Svenson (1981) asked participants to rate their safety and skill as drivers compared to other participants in the experiment. Results showed that the majority of participants rated themselves far above the median, despite the statistical impossibility of more than 50% to be above the median. Here, we embarked on a preregistered very close replication and extension of the Svenson study to examine the replicability of the original finding.

The better-than-average effect

The better-than-average effect has been demonstrated in a variety of domains and is generally considered a manifestation of self-evaluation bias. Drivers believe that they are better drivers (Svenson, 1981; inspired by Preston and Harris, 1965), college instructors believe they are better teachers (Cross, 1977), social psychologists believe they are better researchers (Van Lange et al., 1997), couples believe they have better marriages (Rusbult et al., 2000), and undergradu-

ates believe they have better leadership skills, athletic prowess, and ability to get along with others (Brown, 1986). People even believe that they are less biased than others, an effect known as the *bias blind spot* (Pronin et al., 2002). A recent meta-analysis of 124 published articles found that the better-than-average effect was large and robust across studies (Zell et al., 2020). Although it is closely related to several other biases, including *unrealistic optimism* (predicting that positive outcomes are more likely and negative outcomes are less likely to happen to oneself compared to others; Shepperd et al., 2013; Weinstein, 1980) and the Dunning-Kruger effect (overestimating the rank of one's performance compared to objective measures; Dunning, 2011), the better-than-average effect is unique in that it involves comparing the present self to an average other on a relatively enduring attribute or skill.

Much research has been dedicated to finding boundary conditions and explanations for the effect (for reviews, see Alicke and Govorun, 2005; Chambers and Windschitl, 2004; Moore and Healy, 2008; Sedikides and Alicke, 2012, 2019; Sedikides and Gregg, 2008; Zell et al., 2020). Yet, to the best of our knowledge, no direct replications exist of the original finding by Svenson (1981). The importance of replicability has received increasing recognition in the field of psychological science over the past few years (e.g., Asendorpf et al.,

2013; Brandt et al., 2014; Camerer et al., 2018; Nosek and Errington, 2020; Nosek et al., 2021; Open Science Collaboration, 2015; Zwaan et al., 2017). Replication is considered a cornerstone of science, yet it is only recently that researchers have begun to systematically investigate the replicability of published findings. We here revisit the classic phenomenon to examine the replicability of the original finding with an independent replication.

Choice of study for replication

We chose the Svenson (1981) study based on two factors: absence of direct replications and impact. To the best of our knowledge, there are no published direct replications of this study 1 thus far.[1] The article has had significant impact on scholarly research in several areas of psychology, including social psychology and judgment and decision making. At the time of writing, there were 2,112 citations of the article in Google Scholar.

Findings in the original article

In the original study by Svenson (1981), participants were asked to rate either their skill or their safety as drivers in relation to other participants in the experiment. Data was collected in Sweden (n = 80) and in the US (n = 81) in lab experiments. The results indicated that the majority of participants regarded themselves as more skillful and less risky than the average driver in each group respectively. Among the Swedish participants, 77% ranked their safety as above average and 69% ranked their skill as above average. Among the US participants, 88% ranked their safety as above average and 93% ranked their skill as above average.

Adjustments and extensions

We had to make several adjustments to the original design. First, rather than including two different samples for the two different questions, we ran the questions together in a within- subjects design that would allow us to compare the effects of the two questions and their associated dependent variables. Second, we had to adjust the questionnaire to match the target sample—online American Amazon Mechanical Turk (MTurk) workers. We first introduced a few verification questions to ensure that workers were drivers. Because our study was conducted online, we also had to adjust the reference group. We chose to focus on the US state as the reference group. Third, we had issues with reproducing the question used to elicit rankings and so had to make adjustments. When doing so, we noticed issues with the 10 categories used for percentile ranks (e.g., the midpoint is grouped as 41–50%, and the first

category includes a range of 11 percentiles compared to other categories with a range of 10). We therefore added an extension and chose to randomize the dependent variable question across three designs: (1) our best estimate of what the target article used, (2) an adjusted 11-item scale with a mid-point indicated as 50% (average), and (3) a simple 7-item Likert scale asking participants to compare to the average. We compared effects across the three designs. Thus, the use of three different response scales helps to check the robustness of the effect, as minor methodological features can influence the results of hypothesis tests (e.g., Baribault et al., 2018; Landy et al., 2020).

Method

We report all measures, conditions, data exclusions, and how we determined sample size.

Participants

A total of 1,203 American Amazon Mechanical Turk (MTurk) participants completed the study using TurkPrime.com (M_{age} = 40.40, SD = 12.21; 641 females). A comparison of the target article sample and the replication samples is provided in Table 1. An a priori power analysis in G*Power 3.1 (exact test, two-tailed, with 95% power) indicated that 90 participants were needed to obtain the smallest effect size from the original paper, Cohen's g = 0.19 (see Supplementary Materials). However, a sample size of n = 90 is smaller than the sample size in the original study (n = 161) and is based on an effect size estimate that might be larger than the true effect size. Therefore, we decided to follow suggestions from Simonsohn (2015) and aim for 2.5 times the original sample size. The data collection was combined with data collection for a different study (see Chen et al., 2021, Experiment 2) that required a much larger sample size (studies displayed in randomized order).

Participants first consented to participate in the study and were then asked verification questions regarding having a driver's license, year and location of license,

---

[1]At the time of writing, a Google Scholar search for the term "replication" within works citing Svenson (1981) yielded 259 results, but of these, we found none that would count as a very close replication in the framework of LeBel et al. (2017) while also having high statistical power. There are some studies that essentially replicate the original study (Groeger and Brown, 1989; Svenson et al., 1985); however, sample sizes are relatively small and there are methodological differences especially in terms of the response scale format. In sum, although not a perfect method, we take the results of our search as a strong likelihood that no close replication has been conducted.

Table 1

Differences and similarities between samples in original study and replication

	Svenson (1981)	Replication
Sample size	161	1,203
Geographic origin	US American & Sweden	US American
Gender	Unknown	562 males, 641 females
Median age (years)	22 (US), 33 (Sweden)	37
Average age (years)	Unknown	40.40
Age range (years)	Unknown	18–87
Medium (location)	Lab (Sweden & US)	Computer (online)
Compensation	Unknown	Nominal payment
Year	Before 1981	2019

and state of residence. Participants ($n = 84$) who indicated they did not have a drivers' license were filtered out.

Procedure

Participants indicated how safe and how skilled they were as drivers (both questions included, displayed in random order). They then answered a funneling section and provided demographic information (age, gender, country of birth, family social class, English understanding of study), before being debriefed.

Measures

There were two dependent variables: driving safety and driving skill. The question about safety was phrased as follows:

> We would like to know what you think about how safely you drive an automobile. All drivers are not equally safe drivers. We want you to compare your own skill to the skills of other people in your state. By definition, there is a least safe and a most safe driver. We want you to indicate your own estimated position among the people in your state. Of course, this is a difficult question because you do not know all the people in your state, much less how safely they drive. But please make the most accurate estimate you can.

The question about skill was phrased as follows:

> We would like to know what you think about how skilled you are at driving an automobile. All drivers are not equally skilled drivers. We want you to compare your own skill to the skills of other people in your state. By definition, there is a least skilled and a most skilled driver. We want you to

indicate your own estimated position among the people in your state. Of course, this is a difficult question because you do not know all the people in your state, much less how skilled drivers they are. But please make the most accurate estimate you can.

For each question, participants indicated their driving safety/skill compared to the average driver in their state using one of the three following response scales:

1. **Reproduced materials**: "Please indicate how [safely you drive/skilled you are] compared to others by marking your estimated position among drivers in your state" in 10 categories from 0–10% (*least safe/skilled drivers*) [...] 41-50%, 51-60%, [...] 91–100% (*most safe/skilled drivers*).

2. **11-item scale to include midpoint (changes underlined)**: Same question as above but with the following scale: 0–9% (*least safe/skilled drivers*) [...] 40–49%, 50% (average), 51-60%, [...] 91-100% (*most safe/skilled drivers*).

3. **Standard comparison 7-item Likert scale**: "Please indicate how [safely you drive/skilled you are] compared to others by marking your estimated position compared to other drivers in your state" (1 = *far below average*; 4 = *average*; 7 = *far above average*).

Evaluation criteria for replication

Table 2 provides a classification of the replication using criteria by LeBel et al. (2017). We summarize the replication as a "very close replication". We compare the replication effects with the original effects in the target article using criteria from LeBel et al. (2019).

Table 2

Classification of the replication, based on LeBel et al. (2017)

Design facet	Replication
IV operationalization	Same
DV operationalization	Same
IV stimuli	Same
DV stimuli	Same
Procedural details	Different
Physical settings	Different
Contextual variables	Different
Replication classification	Very close replication

Data analysis

The original article did not include any statistical tests, and the scale and design make it difficult to conduct such a test. Yet, our best estimation of an analysis is to compare the percentages of participants who answered the 50%+ categories and compare those to an expected 50% (binomial test). For the Likert scale, we conducted a one-sample t-test comparing to the mean of 4, the scale midpoint. We examined normality in the distribution of frequencies, including parameters of skewness and kurtosis. Analysis code can be found in the supplementary materials.

Results

Replication

Descriptive statistics of all measures are presented in Table 3. Statistical tests of the hypotheses are summarized in Tables 4–5 and plotted in Figures 1–3.

The medians for the distributions of safety judgments in Table 3 fall in the interval 71–80%, for both percentile category response scales. This indicates that half of the participants believed themselves to be among the safest 30 percent of drivers. Over 90% of participants (93% for the reproduced materials and 91% for the adjusted materials with a 50% midpoint) believed themselves to be safer than the median driver. Binomial tests against test proportion 0.50 (two-tailed) indicated that this effect was statistically significant, $ps < .001$ (see Table 4). In comparison, the original study found that the medians for the distributions of safety judgments fell in the interval 81–90% for the US group and 71–80% for the Swedish group, indicating that half of the participants believed themselves to be among the safest 20 (US) or 30 (Sweden) percent of the drivers in the two groups respectively. 88% in the US group and 77% in the Swedish group believed themselves to be safer than the median driver.

The medians for the distributions of skill judgments in Table 3 fall in the interval 71–80% (both for the reproduced and for the adjusted materials). This indicates that half of the participants believed themselves to be among the most skilled 30 percent of drivers. 91% (for the reproduced materials) and 78% (for the adjusted materials) believed themselves to be more skilled than the median driver. Binomial tests against test proportion 0.50 (two-tailed) indicated that this effect was statistically significant, $ps < .001$ (see Table 4). In comparison, the original study found that the medians for the distributions of skill judgments fell in the interval 61-70% for the US group and 51-60% for the Swedish group. 93% in the US sample and 69% in the Swedish sample believed themselves to be more skilled than the median driver.

When participants rated themselves on a 7-item Likert response scale, they also rated themselves as significantly safer than average, $M = 5.50$ ($SD = 1.08$), $t(386) = 27.28$, $p < .001$, $g = 1.39$, 95% CI [1.25, 1.53], and more skilled than average, $M = 5.28$ ($SD = 1.08$), $t(378) = 23.13$, $p < .001$, $g = 1.19$, 95% CI [1.06, 1.32] (see Table 5).

Extensions

Figure 4 shows the effect size (Hedges's g) and 95% confidence intervals for each rating scale. For skills ratings, the CIs are overlapping in all cases, suggesting no evidence for a difference in the size of the better-than-average effect depending on the type of rating scale used. For safety ratings, the CIs for the two scales that involve percentile categories are overlapping, but the CIs for the Likert scale are slightly lower, suggesting a slightly smaller better-than- average effect when safety is rated on a Likert scale. Nevertheless, the effect is very large in all cases. For effect sizes, confidence intervals, and important study characteristics of the replication, original study, and meta-analysis by Zell et al. (2020), see Supplementary Table S7.

Figure 5 shows the mean safety and skills ratings in each state (excluding states with fewer than 5 responses). We find no obvious pattern in the effect across states. However, some states had very few observations and CIs are generally very large, which complicates interpretation. Therefore, we chose not to analyze this data further.

Exploratory analyses (not pre-registered)

A series of exploratory OLS regressions were run to investigate whether participants' gender, age, and driving experience (i.e., years since driver's license was obtained) predicted their ratings of driving safety and skill. The regressions also included item order (i.e.

Table 3

Proportion of participants in each category

Panel A: Percentile categories used in original study

	N	0-10	11-20	21-30	31-40	41-50	51-60	61–70	71-80	81-90	91-100	
Safety	413	0.2%	0.0%	0.5%	1.9%	3.9%	7.5%	15.3%	24.2%	28.3%	18.2%	
Skill	405	0.2%	0.0%	0.5%	2.2%	5.7%	14.6%	16.0%	26.2%	21.7%	12.8%	

Panel B: 10-percentile categories with 50% midpoint

	N	0-9	10-19	20-29	30-39	40-49	50	51-60	61-70	71-80	81-90	91-100
Safety	403	0.0%	0.2%	0.5%	1.2%	1.5%	6.0%	8.2%	13.2%	26.6%	29.5%	13.2%
Skill	419	0.0%	0.2%	1.2%	2.9%	2.9%	14.6%	10.0%	16.2%	25.5%	18.4%	8.1%

Panel C: Likert response scale

	N	1	2	3	4	5	6	7				
Safety	387	0.5%	2.6%	1.8%	16.3%	25.3%	38.5%	17.3%				
Skill	379	0.0%	1.3%	3.2%	20.6%	25.9%	39.3%	9.8%				

Table 4

Summary of statistical tests for the items with percentile categories

	Category	*N*	Observed prop.	[95% CI]	Test prop.	*p*	Interpretation
			Percentile categories used in original study				
Safety	>50	386	0.94	[0.91, 0.96]	0.50	<.001	Signal – consistent
	<50	27	0.06				
Skill	>50	370	0.91	[0.88, 0.94]	0.50	<.001	Signal – consistent
	<50	35	0.09				
			10-percentile categories with 50% midpoint				
Safety	>50	365	0.91	[0.87, 0.93]	0.50	<.001	Signal – consistent
	<50	38	0.09				
Skill	>50	328	0.78	[0.74, 0.82]	0.50	<.001	Signal – consistent
	<50	91	0.22				

Note. Binomial tests comparing the percentage of participants who rated their driving safety and skill as above average to an expected 50%.

whether participants rated safety or skills first) and study order (i.e., whether participants completed this study or the study reported in Chen et al., 2021 first). The analyses revealed that age and driving experience were associated with both safety and skills ratings, such that the rating increased with increasing age and experience (see Supplementary Tables S1–S6). In addition, there was a significant link between gender and safety ratings using the Likert scale and between gender and skills ratings using the Likert scale and the adjusted materials, indicating that women rated themselves lower. However, there was no such link in the other scales; thus, the results involving gender seem to depend on the response scale format and item content. Including item order and study order in the regression analyses did not alter the interpretation of the effects of gender, age, and driving experience. Item order and study order also had no consistent effect on participants' ratings, although completing the Svenson (1981) replication first was associated with higher safety ratings in

one of the scales (the reproduced materials) and rating safety before skills was associated with lower safety ratings in another (the Likert scale; see Supplementary Tables S1–S6). Nevertheless, the regression results address the question of whether gender, age, and driving experience are associated with participants' *ratings* of driving safety and skill; they do not address the question of whether gender, age, and driving experience affect whether participants rate themselves above average. Because the vast majority of participants rated themselves as above average, we did not conduct such an analysis.

Finally, we investigated the correlation between skills and safety ratings in the three response scales. This analysis indicated that participants' skills ratings were positively correlated with their safety ratings in all three scales (original scale: tau = .52, p < .001, n = 122; adjusted scale with 50% midpoint: tau = .48, p < .001, n = 136; Likert scale: tau = .47, p < .001, n = 121).

Table 5

Summary of statistical tests for the Likert scale

	t	*df*	*p*	Mean diff [95% CI]	Hedges's *g* [95% CI]	Interpretation
Safety	27.38	386	<.001	1.50 [1.40, 1.61]	1.39 [1.25, 1.53]	Signal – consistent
Skill	23.13	378	<.001	1.28 [1.17, 1.39]	1.19 [1.06, 1.32]	Signal – consistent

Note. One-sample *t*-test, test value: 4.

Figure 1

Proportion of participants in each percentile category of safety ratings and skills ratings, using the same percentile categories as the original article.
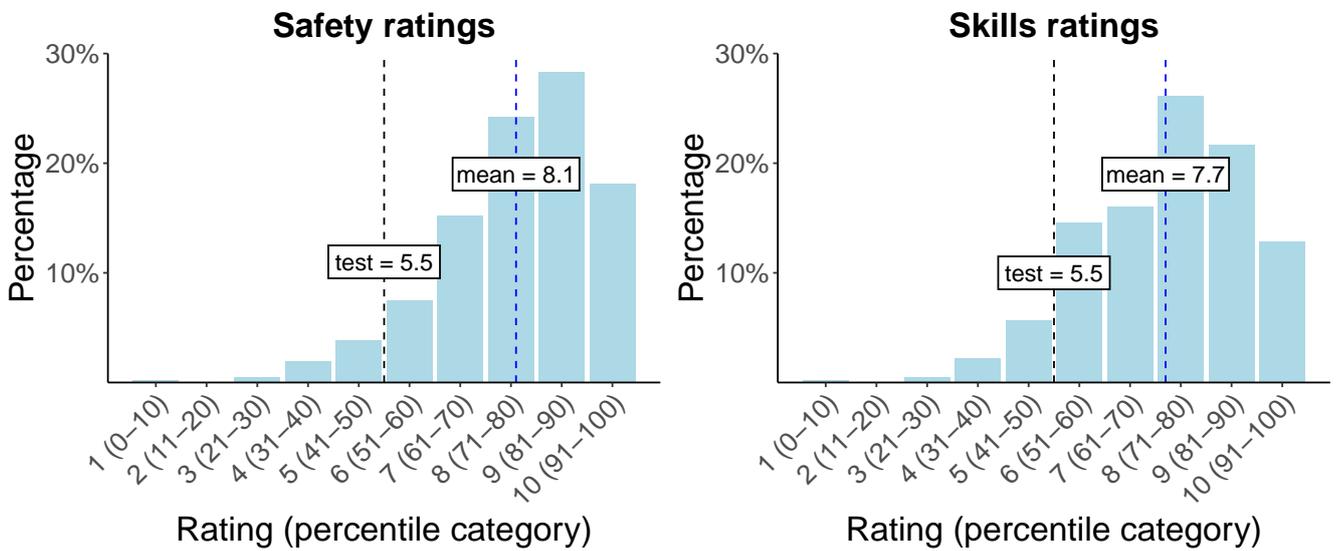
Figure 2

Proportion of participants in each percentile category of safety ratings and skills ratings, using the adjusted percentile categories.
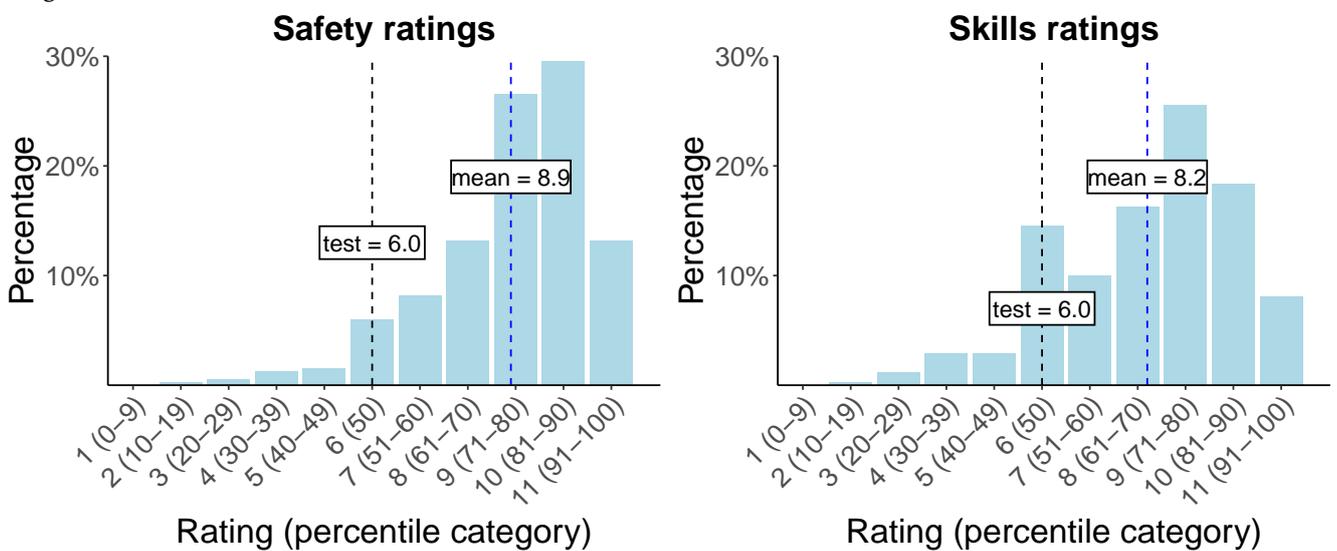
Figure 3

Proportion of participants in each percentile category of safety ratings and skills ratings, using the Likert scale.
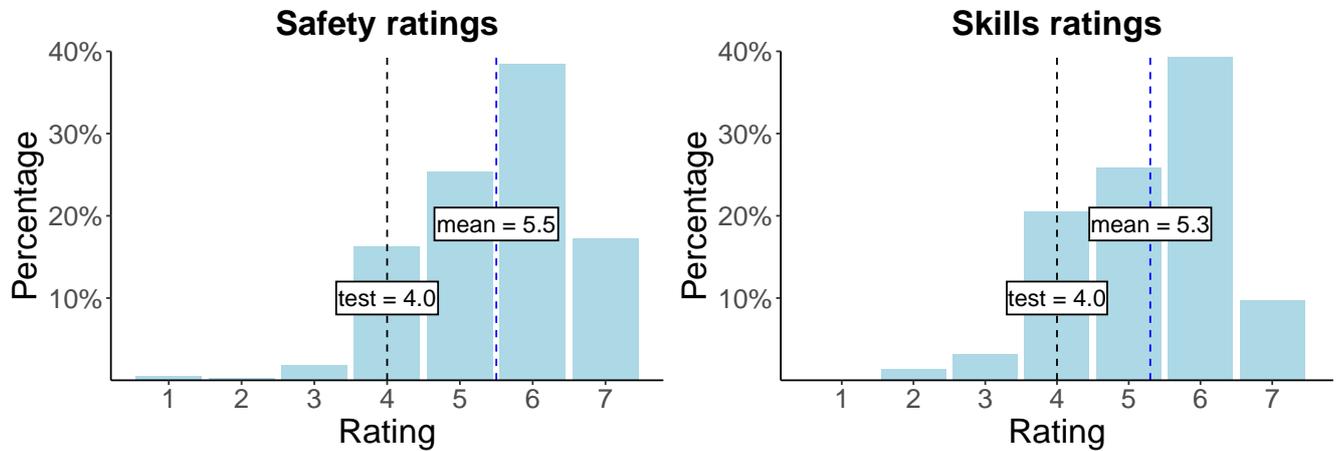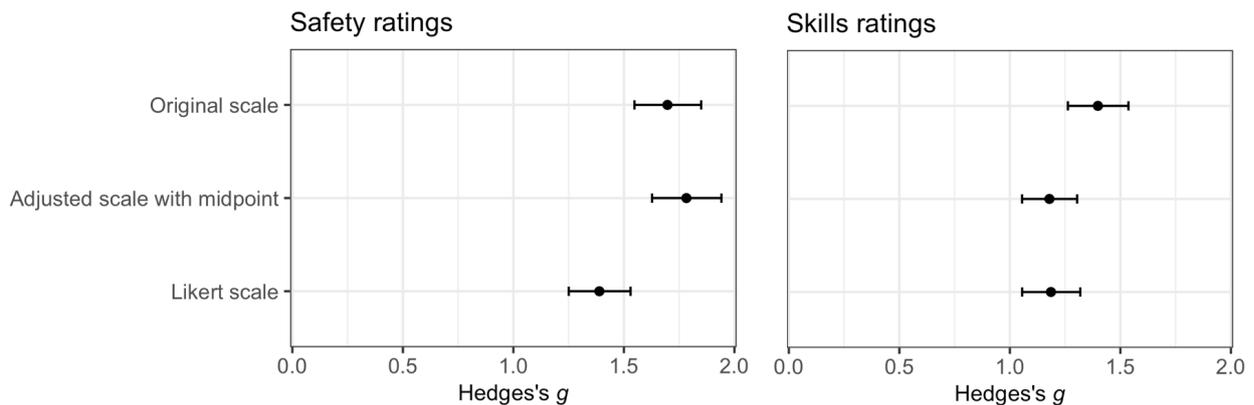

Figure 4

*Effect sizes (Hedges's g) and 95% CIs for each rating scale.*

Discussion

We embarked on a preregistered replication and extension of a classic phenomenon in the judgment and decision-making literature known as the better-than-average effect. The original article found that the majority of participants reported that they were safer and more skilled than the average driver (Svenson, 1981). The findings from our replication are consistent with the original findings. That is, the majority of participants rated their driving safety and skill as above the median. Results were stable across three different response scales: our best estimate of the original materials, an adjusted scale with a 50% midpoint, and a 7-item Likert scale.

Our replication adds to a larger literature investigating the replicability of published research in psycholog-ical science (e.g., Camerer et al., 2018; Open Science Collaboration, 2015). Importantly, our study design closely follows the original study by Svenson (1981) and thereby classifies as a *very close* replication according to replication criteria by LeBel et al. (2017). Recently, Ziano et al., 2020 conducted a replication of another classic study on the better-than- average effect (Alicke, 1985), which indicated that college students' ratings of how characteristic a trait was of them (vs. an average student) increased with increasing desirability of the trait, and that this effect was stronger among more controllable traits. Findings from Ziano et al. (2020) were consistent with the original findings. In sum, findings from the present study are in line with the view of the better-than-average effect as a robust phenomenon.
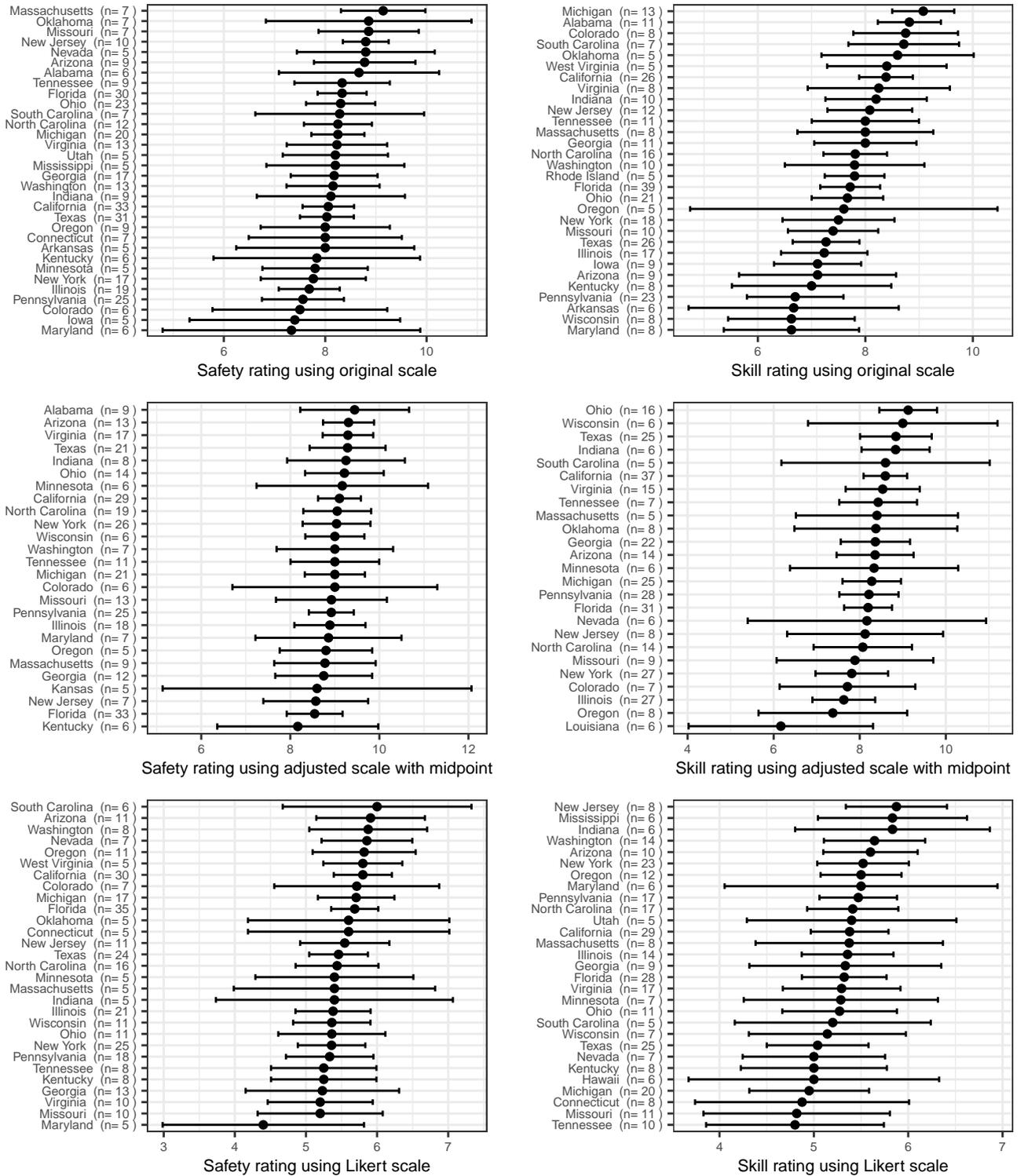
Figure 5

Mean safety and skills ratings in each state (excluding states with fewer than 5 observations). Error bars represent 95% CIs.

Author Contact

Lina Koppel, ORCID: 0000-0002-6302-0047. Gustav Tinghög, ORCID: 0000-0002-8159-1249. Daniel Västfjäll, ORCID: 0000-0003-2873-4500. Correspondence: Gilad Feldman, Department of Psychology, University of Hong Kong, Hong Kong SAR; gfeldman@hku.hk; ORCID: 0000-0003-2812-6599

Conflict of Interest and Funding

Author Contributions

Role	LK	DA	GT	DV	GF
Conceptualization	X	X	X	X	X
Pre-registration	X	X	X	X	X
Data curation					X
Formal analysis	X	X			X
Funding acquisition					X
Investigation	X	X	X	X	X
Pre-registration peer review/verification	X	X	X	X	X
Data analysis peer review/verification	X	X	X	X	X
Methodology	X	X	X	X	X
Project administration	X				X
Resources					X
Software	X	X	X	X	X
Supervision					X
Validation	X	X	X	X	X
Visualization	X	X			X
Writing – original draft	X				
Writing – review and editing	X	X	X	X	X

Target article

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? Acta Psychologica, 47, 143-148. https://doi.org/10.1016/0001-6918(81)90005-6.

Links to project files

Project page on OSF with datasets and code: https://osf.io/fxpwb/. Pre-registration (including materials and analysis code): https://osf.io/jky24.

Acknowledgments

Open Science Practices

This article earned the Preregistration+, Open Data and the Open Materials badge for preregistering the hypothesis and analysis before data collection, and for making the data and materials openly available. It has been verified that the analysis reproduced the results presented in the article. The entire editorial process, including the open reviews, are published in the online supplement.

References

Alicke, M. D. (1985). Global self-evaluation as determined by the desirability and controllability of trait adjectives. *Journal of Personality and Social Psychology*, *49*(6), 1621–1630. https://doi.org/10.1037/0022-3514.49.6.1621

Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. In M. D. Alicke, D. A. Dunning, & L. E. Krueger (Eds.), *The self in social judgment* (pp. 85–106). Psychology Press.

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., Van Aken, M. A., Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., Van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (2018). Metastudies for robust tests of theory. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2607–2612. https://doi.org/10.1073/pnas.1708285114

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*(1), 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

Brown, I. D. (1986). Evaluations of self and others: Self-enhancement biases in social judgments. *Social Cognition*, *4*(4), 353–376. https://doi.org/10.1521/soco.1986.4.4.353

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Chambers, J. R., & Windschitl, P. D. (2004). Biases in social comparative judgments: The role of nonmotivated factors in above-average and comparative-optimism effects. *Psychological Bulletin*, *130*(5), 813–838. https://doi.org/10.1037/0033-2909.130.5.813

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. (2021). 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

Cross, K. P. (1977). Not can, but will college teaching be improved? *New Directions for Higher Education*, *17*, 1–15. https://doi.org/10.1002/he.36919771703

Dunning, D. (2011). The dunning-kruger effect: On being ignorant of one's own ignorance. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (pp. 247–296). Academic Press. https://doi.org/10.1016/B978-0-12-385522-0.00005-6

Groeger, J., & Brown, I. (1989). Assessing one's own and others' driving ability: Influences of sex, age, and experience. *Accident Analysis Prevention*, *21*(2), 155–168. https://doi.org/10.1016/0001-4575(89)90083-3

Landy, J., Jia, M., Ding, I., Viganola, D., Tierney, W., Dreber, A., Johanneson, M., Pfeiffer, T., Ebersole, C., Gronau, Q., Ly, A., van den Bergh, D., Marsman, M., Derks, K., Wagenmakers, E.-J., Proctor, A., Bartels, D. M., Bauman, C. W., Brady, W. J., . . . Uhlmann, E. L. (2020). Crowd-sourcing hypothesis tests: Making transparent how design choices shape research results. *SSRN Electronic Journal*, *146*(5), 451–479. https://doi.org/10.2139/ssrn.3654406

LeBel, E., Berger, D., Campbell, L., & Loving, T. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, *113*(2), 254–261. https://doi.org/10.1037/pspi0000106

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

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–517. https://doi.org/10.1037/0033-295X.115.2.502

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, *18*(3), 1–8. https://doi.org/10.1371/journal.pbio.3000691

Nosek, B. A., Hardwicke, T. E., Corker, K. S., & Rohrer, J. (2021). Replicability , robustness , and reproducibility in psychological science. *Annual Review of Psychology*. https://doi.org/10.31234/osf.io/ksfvq

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Preston, C. E., & Harris, S. (1965). Psychology of drivers in traffic accidents. *Journal of Applied Psychology*, *49*(4), 284–288. https://doi.org/10.1037/h0022453

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. *Personality and Social Psychology Bulletin*, *28*(3), 369–381. https://doi.org/10.1177/0146167202286008

Rusbult, C. E., Van Lange, P. A., Wildschut, T., Yovetich, N. A., & Verette, J. (2000). Perceived superiority in close relationships: Why it exists and persists. *Journal of Personality and Social Psychology*, *79*(4), 521–545. https://doi.org/10.1037/0022-3514.79.4.521

Sedikides, C., & Alicke, M. D. (2012). Self-enhancement and self-protection motives. In R. M. Ryan (Ed.), *Oxford handbook of motivation* (pp. 303–322). Oxford University Press. https://doi.org/10.7873/date.2014.002

Sedikides, C., & Alicke, M. D. (2019). The five pillars of self-enhancement and self-protection. In M. Ryan (Ed.), *The oxford handbook of human motivation* (2nd ed., pp. 307–319). Oxford University Press.

Sedikides, C., & Gregg, A. P. (2008). Self-enhancement: Food for thought. *Perspectives on Psychological Science*, *3*(2), 102–116. https://doi.org/10.1111/j.1745-6916.2008.00068.x

Shepperd, J. A., Klein, W. M., Waters, E. A., & Weinstein, N. D. (2013). Taking stock of unrealistic optimism. *Perspectives on Psychological Science, 8*(4), 395–411. https://doi.org/10.1177/1745691613485247

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica, 47*, 143–148. https://doi.org/10.1016/0001-6918(81)90005-6

Svenson, O., Fischhoff, B., & Macgregor, D. (1985). Perceived driving safety and seatbelt usage. *Accident; analysis and prevention, 17*(2), 119–113. https://doi.org/10.1016/0001-4575(85)90015-6

Van Lange, P. A., Taris, T. W., & Vonk, R. (1997). Dilemmas of academic practice: Perceptions of superiority among social psychologists. *European Journal of Social Psychology, 27*(6), 675–685. https://doi.org/10.1002/(SICI)1099-0992(199711/12)27:6{\textless}675::AID-EJSP838{\textgreater}3.0.CO;2-F

Weinstein, N. D. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology, 39*(5), 806–820. https://doi.org/10.1037/0022-3514.39.5.806

Zell, E., Strickhouser, J. E., Sedikides, C., & Alicke, M. D. (2020). The better-than-average effect in comparative self-evaluation: A comprehensive review and meta-analysis. *Psychological Bulletin, 146*(2), 118–149. https://doi.org/10.1037/bul0000218

Ziano, I., Mok, P. Y., & Feldman, G. (2020). Replication and extension of alicke (1985) better-than-average effect for desirable and controllable traits. *Social Psychological and Personality Science*. https://doi.org/10.1177/1948550620948973

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2017). Making replication mainstream. *Behavioral and Brain Sciences, 41*, 1–50. https://doi.org/10.1017/S0140525X17001972