Title

Looking for evidence of the Dunning-Kruger effect: an analysis of 2400 online test takers

Abstract

The Dunning-Kruger effect is a well-known psychological finding. Unfortunately, there are two aspects of the finding, one trivial and the other an unsupported theory that purports to explain this pattern. Recently, (Gignac & Zajenkowski, 2020) suggested two ways to operationalize and test the proposed theory. They found no support for the theory's predictions. We carried out a replication of their study using archival data from a large dataset of online subjects (n = 2,413). We used two measures of self-estimated ability: estimated sumscore (correct responses), and estimated own-centile. Both had strong correlations with objective performance (r's .50 and .54), but we find no evidence of nonlinearity for either. We find some limited evidence of heteroscedasticity for self-centile estimates, but not raw score estimates. Overall, the evidence was mostly incongruent with Dunning-Kruger theory.

Keywords: Dunning-Kruger effect, regression towards the mean, intelligence, science knowledge, self-estimated intelligence, self-perception, replication

Introduction

The Dunning-Kruger effect is one of the most popular psychological findings. The original study has collected about 6,600 citations on Google Scholar since being published in 1999 (Kruger & Dunning, 1999). The typical Dunning-Kruger pattern is shown in Figure X.

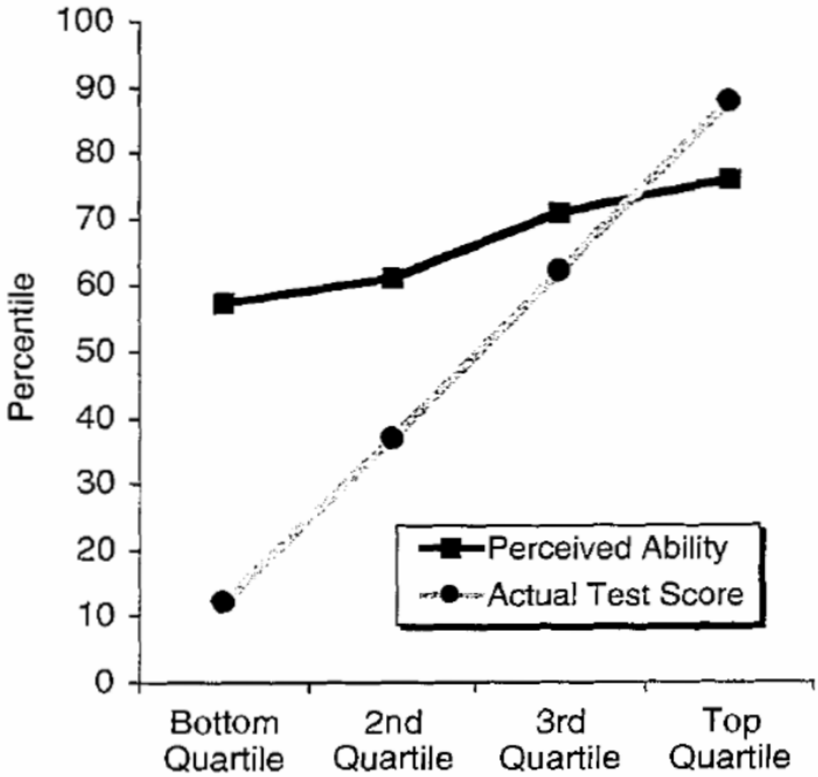

Figure X. Typical Dunning-Kruger pattern. Reproduced from (Kruger & Dunning, 1999).

In the typical Dunning-Kruger pattern, there is a positive relationship between own measure ability and self-estimated ("perceived") ability or performance. However, it can be seen by comparing the two lines that below average persons tend to overestimate themselves quite strongly, while above average persons underestimate themselves, but less strongly than the below average overestimate themselves. The theory advocated to explain this pattern is that below average subjects not only suffer from a lack of ability to perform well, but also suffer from below average metacognitive ability to recognize their own poor performance.

However, criticism of the original study were soon published, though these were mostly ignored (Ackerman et al., 2002; Krueger & Mueller, 2002; see Schimmack, 2020 for a review). In fact, the familiar Dunning-Kruger pattern arises from two simple facts. First, self-estimated ability is positively, but imperfectly, correlated with actual ability. A large meta-analysis found a mean observed r = .33 (Freund & Kasten, 2012). Thus, from a regression perspective, the true ability level of a person is much closer to the mean than a given estimate, which is why this has sometimes also been referring to as an example of regression towards the mean (Krueger & Mueller, 2002; see in general, Dalliard, 2017). Second, there is a general tendency to overestimate own performance in general (called the illusory superiority effect or the better-than-average effect, (Zell et al., 2020)). When these two facts are combined, they yield the familiar Dunning-Kruger pattern, shown in Figure X.

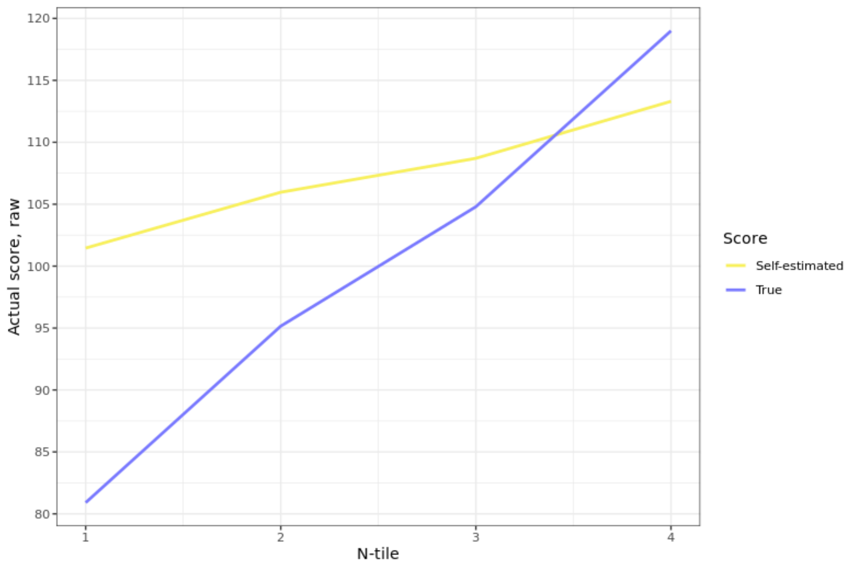

Figure X. An example of the Dunning-Kruger pattern. Based on simulated data from http://emilkirkegaard.dk/understanding statistics/?app=Dunning Kruger. In this case, a correlation of .30 is assumed between true and self-rated scores, as well as a 7.5 IQ (0.50 d) overestimation bias. The values are then binned into 4 bins.

The simulated results closely approximate the real results. Though everybody has some accuracy, the below average subjects are more in error than the above average subjects. Those high in ability tend to underestimate themselves, but less so than the below average subjects overestimate themselves. Because this pattern above arises from two simple statistics facts mentioned above, there is nothing for the meta-cognitive theory advanced by Dunning, Kruger and others to explain, leaving it in an uncertain position.

Recently, however, (Gignac & Zajenkowski, 2020) proposed two different ways to test the theory. The core claim of the Dunning-Kruger theory is that below average subjects on some trait are lacking in a metacognitive ability to estimate themselves correctly in some sense. Their purported evidence is the greater difference between their centile estimates and real estimates. Since this arises trivially from the above two facts, this is not evidence of the Dunning-Kruger theory.[1] However, a different way to operationalize this theory, that is, derive a testable prediction, is that it makes a claim that below average persons should have a weaker association between their self-estimates and the real estimations. In statistical terms, the relationship should exhibit heteroscedasticity with

[1] In the Bayesian sense, to be evidence of something, the data must be more probable on some model compared to another model. However, since the statistical artifact model (overestimation + imperfect linear association) model predicts the exact same Dunning-Kruger pattern, this pattern is not evidence of any Dunning-Kruger model, but must be neutral. Insofar as the simple model involves only known facts and does not require positing a new mechanism, its prior probability is higher is thus the posterior is also higher given the available evidence.

greater residual variance in the below average ability region ('dual burden'). A second derived prediction is that the association between self-estimated and real ability should diverge from the overall trend, with a weaker slope in the below average region. Both of these predictions involve the below average persons being worse in some sense at predicting their own performance level. These predictions are testable using existing methods and data. (Gignac & Zajenkowski, 2020) tested both predictions in a dataset of 929 subjects who had taken the Raven's advanced progressive matrices test (a standard nonverbal intelligence test) and who had estimated themselves on a 1-25 scale. First, they found no evidence of heteroscedasticity using the Glejser test. This test involves saving the residuals from the linear model (*self-estimated ability ~ objectively measured ability*, where ~ denotes "regressed on"), converting to absolute values, and correlating with the predictor (i.e., objectively measured ability). The correlation was -.05 with 95% confidence interval of -.11 to .02. Second, they looked for a nonlinear association using a model comparison with a quadratic model. The model comparison found no incremental validity of the nonlinear model (incremental R^2 < 1%). They plotted the data using a smoothing function (local regression, LOESS), which also showed no notable deviation from linearity. The purpose of this paper was to replicate the findings of (Gignac & Zajenkowski, 2020) in a new and larger sample using more robust methods for testing for heteroscedasticity and nonlinearity.

Data

We used archival data from an online pilot test of a new 25-item science knowledge scale under development (the items can be found in the supplementary materials). During test development, we posted a link to a questionnaire on Twitter with a science knowledge scale. The tweet was retweeted by some prominent users and went viral, resulting in about 2400 subjects taking the test. Aside from their performance on the items, they filled out a few related questions such as their age and educational attainment. Two of these asked them to assess their own performance:

- The previous page featured 25 questions testing your knowledge How many correct answers do you think you gave?
- With regards to your knowledge of science, what percentile of the general population do you think you are in?

All data, R analysis code, and materials are available at https://osf.io/fhqap/. The R notebook can also be viewed at https://rpubs.com/EmilOWK/Dunning_Kruger_2021.

Results

We scored the cognitive ability data using both simple sum scores (sum of correct answers) and item response theory (IRT) analysis (DeMars, 2010), using the 2PL model as implemented in the **mirt** package for R (Chalmers et al., 2020). The latent variable extracted this way we labelled g, since it approximates the general factor of intelligence (Jensen, 1998). Figure X shows the distribution of scores by scoring method. Their correlation was .95. The estimated reliability was .74 for the actual data, and .74 with an assumed perfect normal distribution.[2] These values are probably underestimates of the

[2] For details of the calculation, see documentation for *marginal rxx()* and *empirical rxx()* in the **mirt** package.

test-retest reliability. For instance, the retest reliability for the similar WAIS-R information scale was .81 in a sample of 101 elderly persons who were tested again after 1 year (Snow et al., 1989), and reached .92 in a representative sample of 100 Australians (Shores & Carstairs, 2000). Cronbach's alpha for the same data was .68.

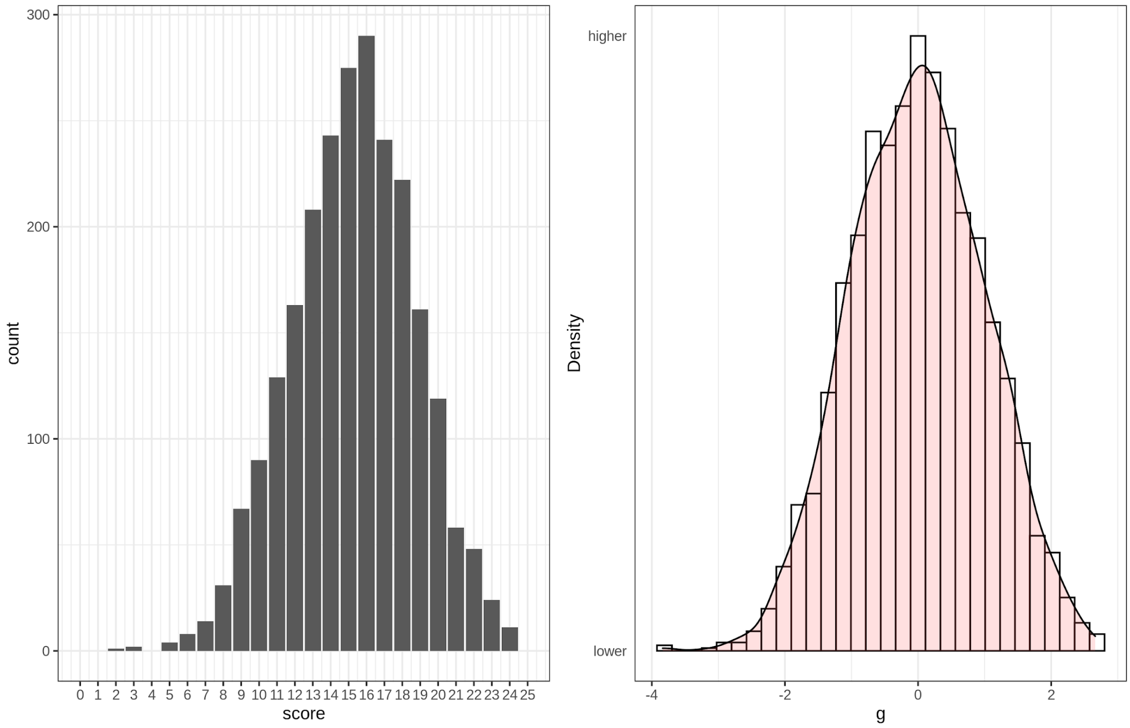

Figure X. Distributions of scientific knowledge by scoring method. Left panel shows sum scores, and the right panel, item response theory standard scores (density curve overlaid).

Both distributions were very normal despite the somewhat unusual recruitment method and the untested test. Since we lack normative data, we don't know which Greenwich IQ (British norms) the results compare to, but this is likely an above average group, as it was recruited from Twitter users' followers who post a lot of science. Table X shows descriptive statistics for the numerical variables.

Var	**n**	**Mean**	**Median**	**SD**	**MAD**	**Min**	**Max**	**Skew**	**Kurtosis**
score	2413	15.28	15.00	3.48	2.97	0.00	25.00	-0.15	0.03
g	2412	0.00	0.00	1.00	1.04	-3.84	2.66	-0.03	-0.23
centile guess	2406	68.26	70.00	20.21	19.27	0.00	100.00	-0.76	0.16
score guess	2408	14.06	14.00	4.29	4.45	0.00	25.00	-0.21	-0.27

age	2412	35.92	33.00	13.44	11.86	12.00	130.00	1.54	5.36
time taken min	2413	20.19	8.85	169.65	3.78	0.37	5619.52	27.39	821.36

Table X. Descriptive statistics for numerical variables. SD = standard deviation, MAD = median absolute deviation from the median. g = general intelligence factor.

The average centile guess was 68th, which is above average (d = 0.47 above the 50th centile). However, since this was a self-selected sample and the question asked about the general population, this estimate is not necessary too high. The mean score guess was 14 and the mean score obtained was 15.3, thus subjects slightly underestimated their ability level. The mean age was 36 with a standard deviation of 13.4. This is somewhat younger than the adult population average (40-45), but much closer to representative than typical college/university student samples. The questionnaire did not ask about sex/gender, so the distribution is unknown. However, based on prior surveys by the author, it is likely to be about 90% male. The median time to take the survey was 8.9 minutes. The educational attainment levels were: 37% bachelor’s, 25% master’s, 12% doctorate. Considering that much of the sample was too young to have completed their education entirely, this is a relatively elite sample.

We had two different sets of self-estimates: one based on the raw score and another based on the centile relative to the general population. Surprisingly, the estimates correlated only at .61, as shown in Figure X. In an ideal world, these two variables should be near perfectly correlated. However, since they were not, this opens questions about differential validity and perhaps incremental validity.

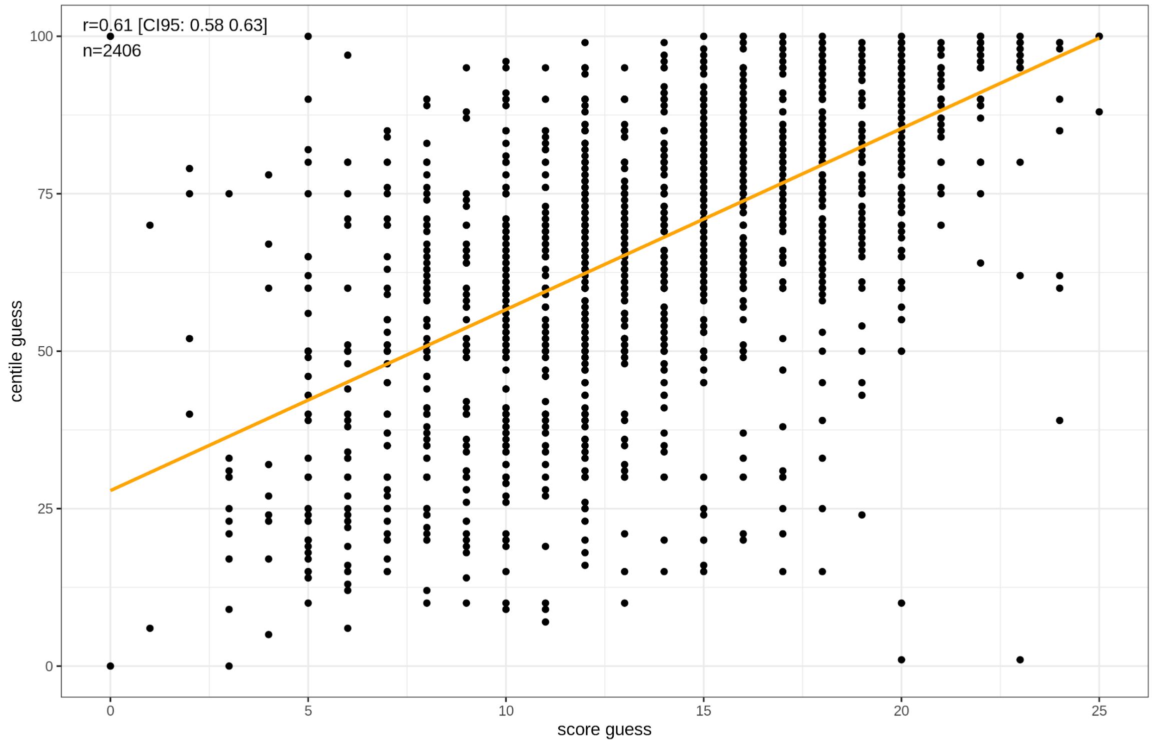

Figure X. Scatterplot of relationship between two different self-estimated ability variables.

The distributions of the self-estimates were otherwise unremarkable, as shown in Figure X.

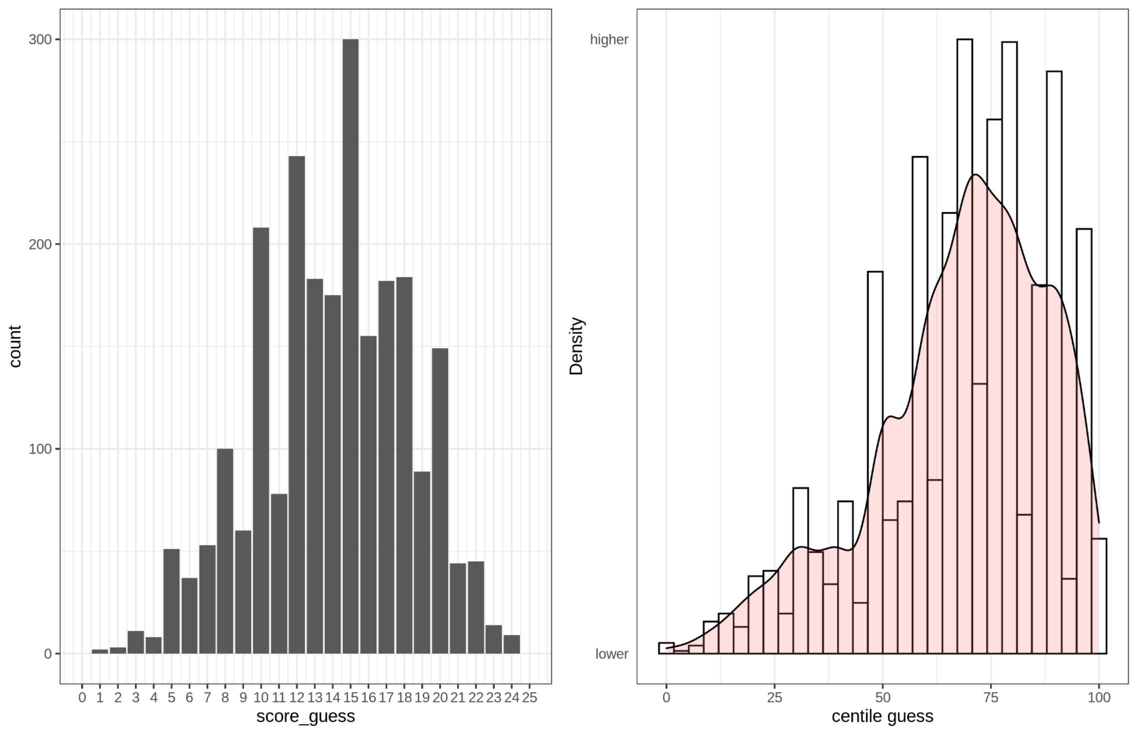

Figure X. Distributions of self-estimated science knowledge.

Moving on to the main tests, since we had two measures of self-estimated ability, we had two main tests of the Dunning-Kruger effect. Figures Xa-b shows the results for the nonlinear fits.

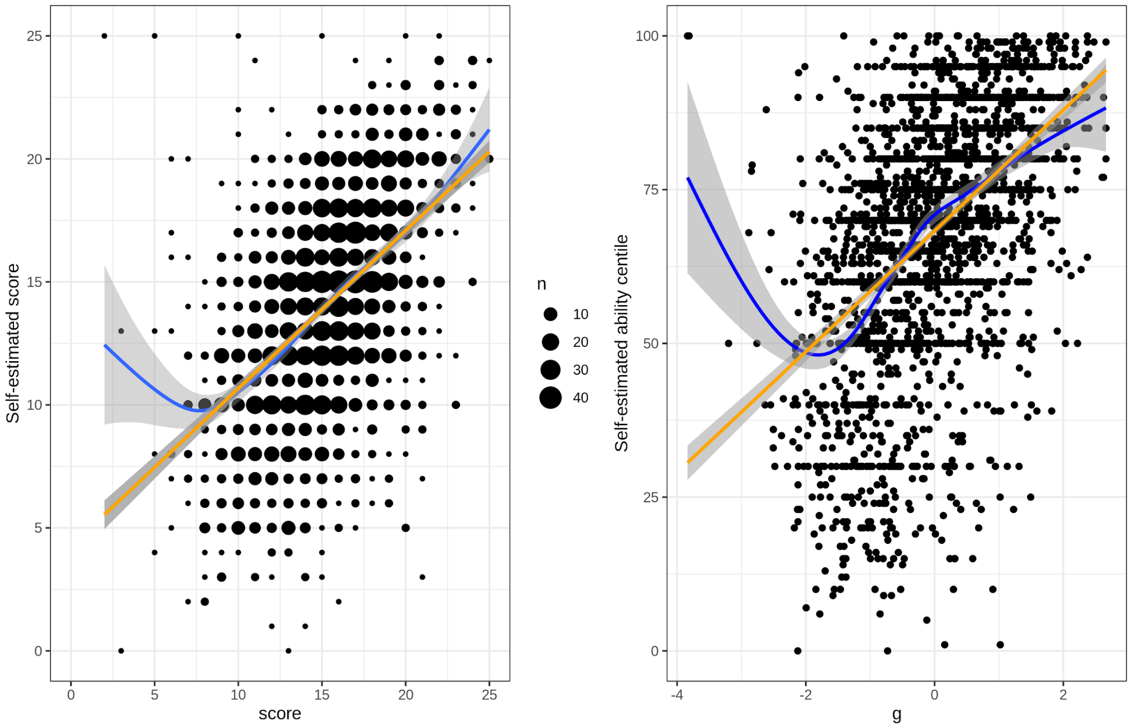

Figure Xa-b. Scatterplots showing linear (orange) and LOESS (blue; locally estimated scatterplot smoothing) fits for self-estimated ability and measured ability. Left plot: sumscore and self-estimated sumscore. Right plot: item response theory g score and self-estimated ability centile.

The correlations are relatively strong: r's .52 and .48, for score ~~ self-estimated score, and g ~~ self-estimated ability centile, respectively (both p's < .001). The relationship between self-estimated ability and objectively measured ability is close to linear with the exception of the area below about -2z. The upward pattern is caused by a few outliers with very poor scores and maximum self-rated ability. These are likely not serious survey responses (Alexander, 2013). We left the outliers in the dataset here to illustrate the dangers of not plotting the data for testing purposes. Figure X shows the same plots but with data points beyond 2.5z in either direction removed (21 cases removed).

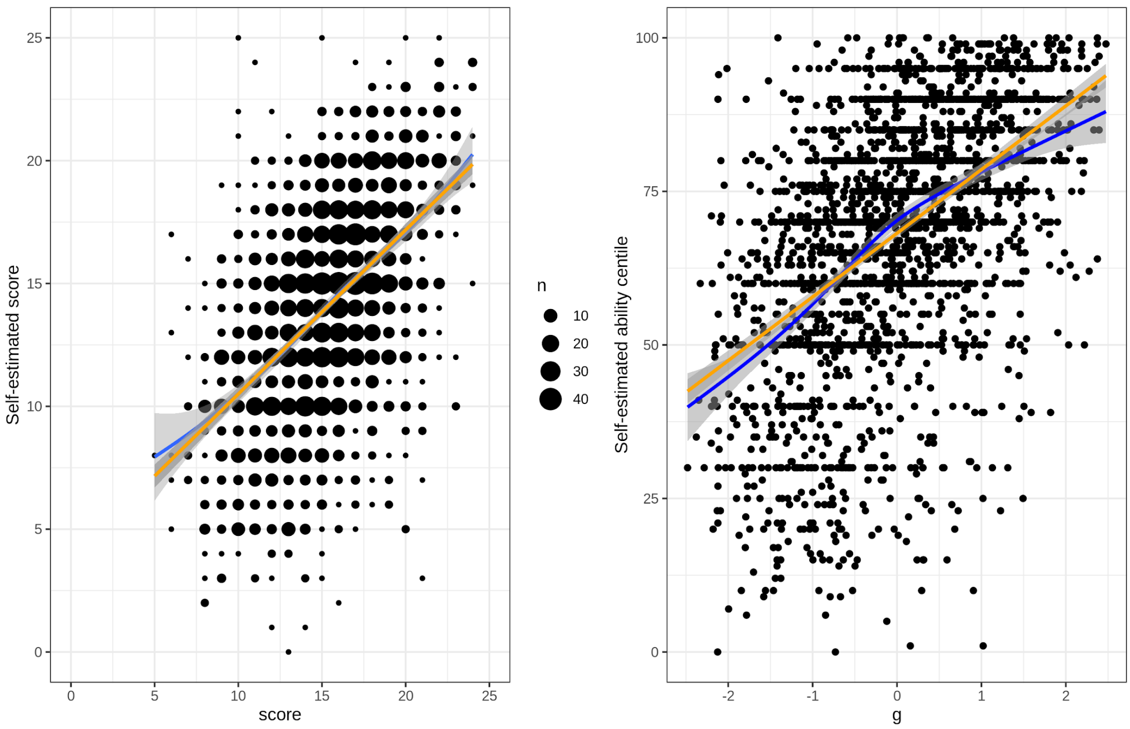

Figure Xa-b. Scatterplots showing linear (orange) and LOESS (blue; locally estimated scatterplot smoothing) fits for self-estimated ability and measured ability. Left plot: sumscore and self-estimated sumscore. Right plot: item response theory g score and self-estimated ability centile. Outliers beyond 2.5 z removed.

We see now in the plots that the association is now near-perfectly linear. To be fair, when tested for nonlinearity using a model comparison (natural spline model vs. linear model) using a likelihood ratio test (specifically, *lrtest()* from the **rms** package, (Harrell, 2019)), one finds small p values (p's .004 and p < .001), and thus evidence for nonlinearity, but the deviation from linearity was very small and not worth caring about (model R2 adj. changes: 0.3% and 1.5%). Thus, contrary to the predictions of the Dunning-Kruger theory, we find essentially no nonlinearity in the estimates.

Turning to the question of heteroscedasticity, we employed the same method as in (Kirkegaard, 2021). The approach is as follows: first, the model of interest is fit. This is the statistical model that wants to evaluate for heteroscedasticity. Second, the residuals are saved, standardized, and then converted to positive (absolute) values. Third, linear and nonlinear models are then fit to the residuals using the predictor of interest to look for evidence of heteroscedasticity. In the simulation study carried out by (Kirkegaard, 2021), it was found that this approach was able to detect real heteroscedasticity, and without excessive false positives. It can also detect the difference between linear and nonlinear heteroscedasticity, though not with optimal statistical properties (elevated false positive rates with regards to confusion between types of heteroscedasticity). Figure X illustrates the concept of heteroscedasticity.

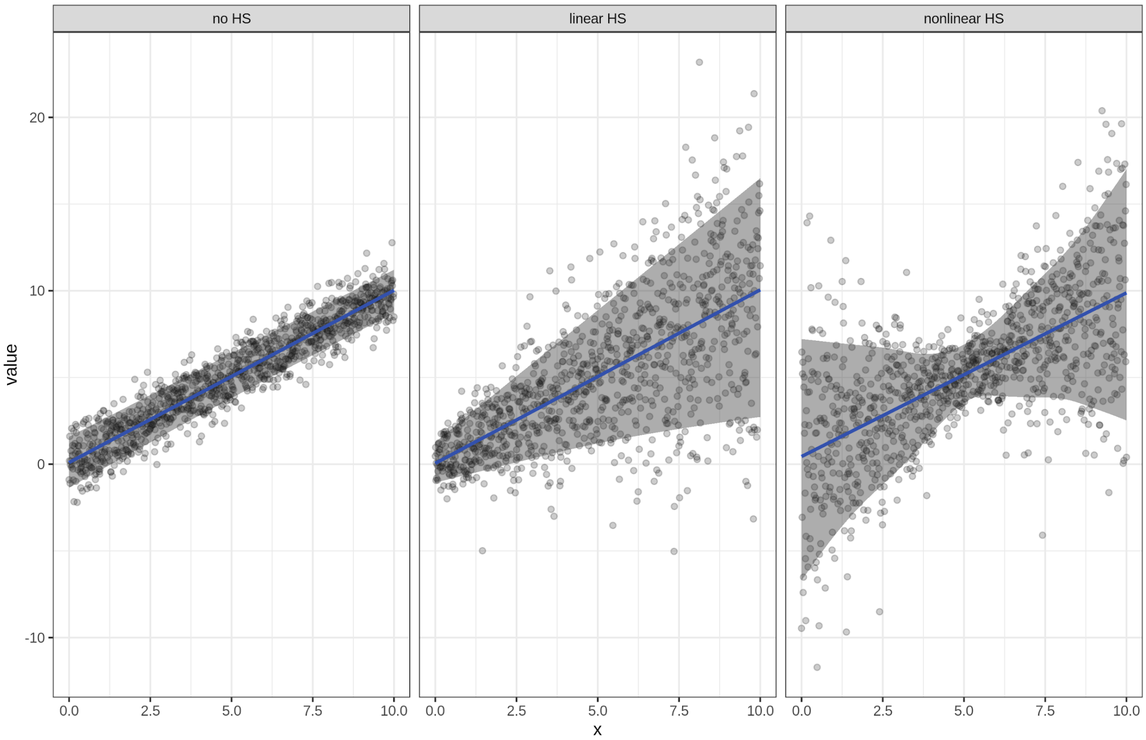

Figure X. Types of heteroscedasticity. Left: no heteroscedasticity (homoscedasticity), 2) monotonic linear increasing heteroscedasticity, and 3) nonmonotonic nonlinear heteroscedasticity.

For the g and centile guess relationship, we find strong evidence of linear heteroscedasticity, and it is concentrated as increased residual variance in the below average region. Figure X shows the estimated 10 and 90th centiles, using quantile general additive model smoothing.

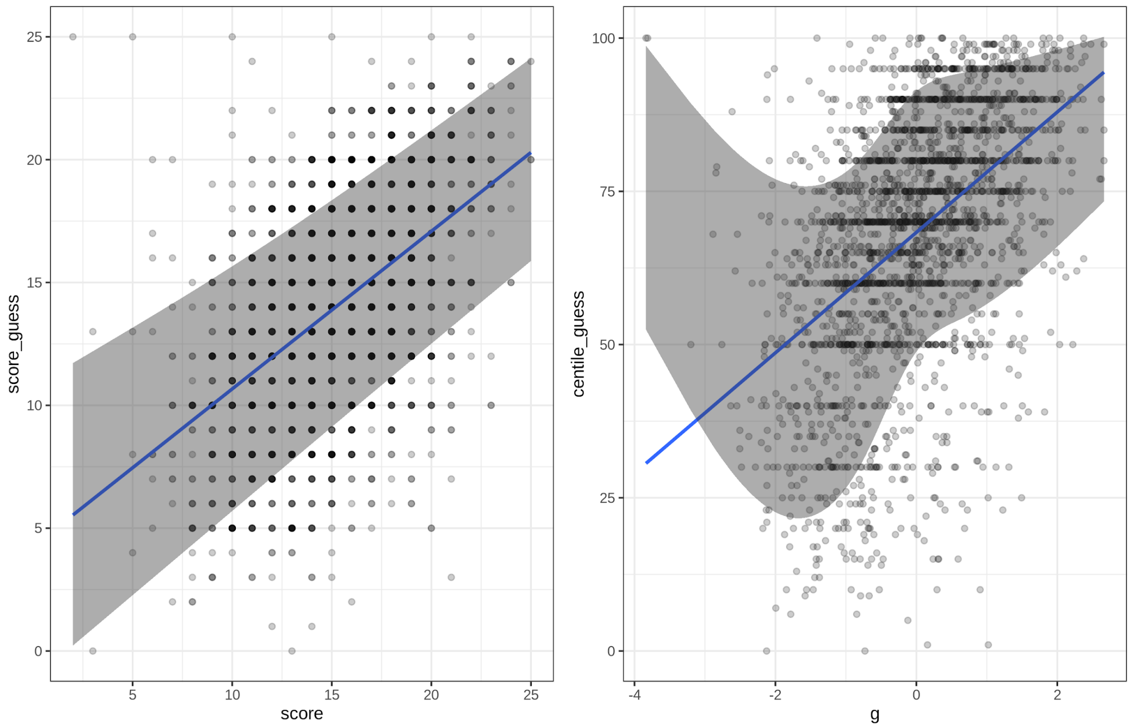

Figure X. Estimated 10th and 90th centiles of ability variables for the complete dataset.

It can be seen that the left plot shows essentially no heteroscedasticity (i.e., the spread around the regression line is constant), while the right plot shows some nonmonotonic nonlinear heteroscedasticity. However, we reasoned this was likely due to the outliers at the very low end of the ability, as was seen in the prior section. Thus, we reran the tests on the reduced dataset. Figure X shows the results.

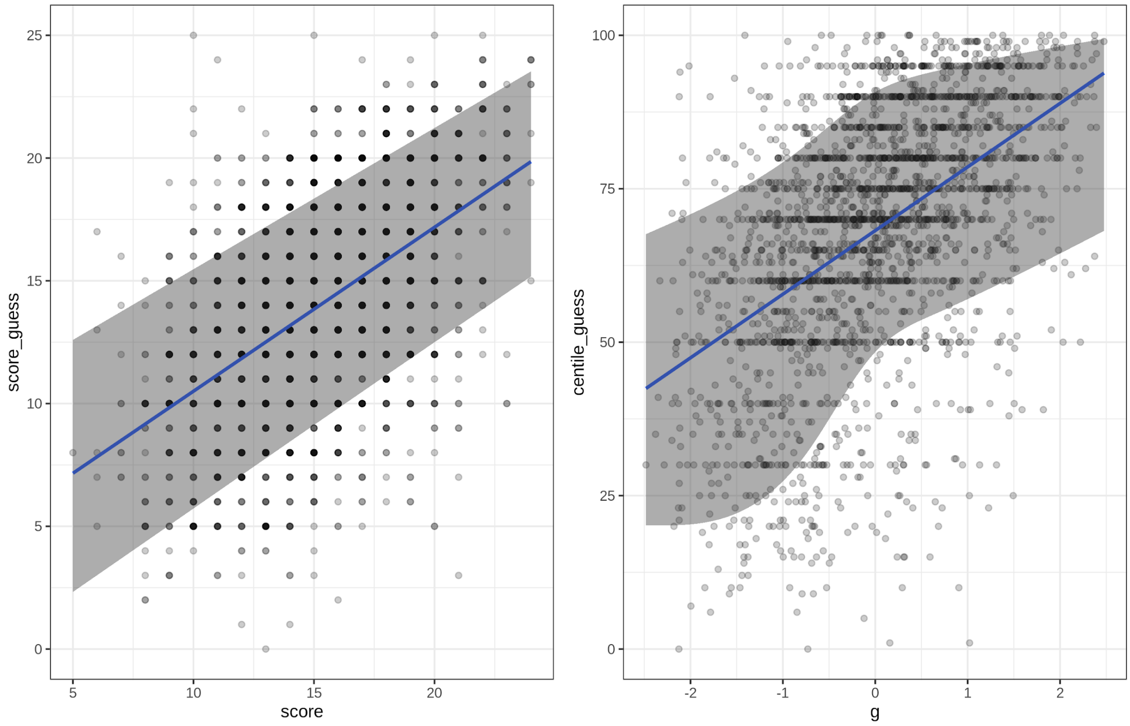

Figure X. Estimated 10th and 90th centiles of ability variables for the dataset with outliers removed.

The pattern for the centiles (right plot) is now simpler. The effect size of the heteroscedasticity seen is not large: about 2% of the variance in the residuals can be explained by the predictor variable ($p < .0001$, linear rank data test). In contrast, the model adj. R2 for the sumscores is 0% ($p = .07$, linear rank data test, left plot). Thus, we find very little heteroscedasticity in the estimates, contrary to predictions of the Dunning-Kruger theory.

Finally, we modeled the data to see if the two ways of measuring self-estimated ability had incremental validity to predict actual ability. Table X shows the model results, while Table X shows the correlation matrix between the ability variables.

	Score	**g**	**Score guess**	**Centile guess**
Score	1.00	0.95	0.53	0.47
g	0.95	1.00	0.54	0.50
Score guess	0.52	0.53	1.00	0.61
Centile guess	0.46	0.48	0.61	1.00

Table X. Correlation matrix. Above diagonal: results based on outlier-filtered dataset. Below diagonal: results based on all datapoints. All correlations $p < .0001$.

Predictor	Sumscore + raw guess	Sumscore + centile guess	Sumscore combined	g + centile guess	g + raw guess	g combined
Intercept	0.00 (0.017)	0.00 (0.018)	0.00 (0.017)	0.00 (0.018)	0.00 (0.017)	0.00 (0.017)
score guess	0.53 (0.017***)		0.38 (0.021***)		0.54 (0.017***)	0.38 (0.021***)
centile guess		0.47 (0.018***)	0.23 (0.021***)	0.49 (0.018***)		0.26 (0.021***)
R2 adj.	0.279	0.221	0.313	0.246	0.295	0.337
N	2388	2386	2386	2386	2388	2386

Table X. Regression model results for incremental validity tests. *** = p < .001. Outcome variable in the 3 leftmost models = sumscore, outcome variable in 3 rightmost models = item response theory g factor score.

The model results show that in each case, there is notable incremental validity in using more than one measure of self-estimated ability. The best predictor to predict own sumscore was the estimated sumscore, but adding the estimated centile added another 4% variance. Results were similar for the g factor scores.

Discussion

We carried out a large replication study looking for evidence of the Dunning-Kruger metacognitive theory using two methods. We found no convincing evidence in favor of the theory. What little is found seems to be mostly due to a small set of outliers who likely provided dishonest data. Our replication study was about 2.5 times larger than the prior study by (Gignac & Zajenkowski, 2020, n=929). For this reason, we have more statistical precision and our study carries more weight. However, both studies found the same result, namely that the patterns the Dunning-Kruger theory should generate were not found. Thus, we successfully replicated (Gignac & Zajenkowski, 2020). Our study thus casts further doubt on the Dunning-Kruger metacognitive theory.

We found substantially stronger correlations between self-estimated ability and measured ability than indicated by a prior meta-analysis on this topic, which found r = .33 (Freund & Kasten, 2012), while we find r’s .50 and .54 (outliers removed). This meta-analysis did not account for known differences in the reliability of ability measures, or range restriction, as the authors were not able to find these in the published studies. The present study used a relatively brief cognitive test (25 items), so it seems unlikely our measure was overall more reliable than their average test (reliability was estimated at .74). Thus, we are not sure why we find a substantially higher correlation. The moderator analysis in the meta-analysis found that using relative self-ratings (such as centiles; estimated r = .33 + .09 = .42) produced stronger correlations, whereas we find the raw score estimate produced slightly stronger correlations. They did not study self-estimates on knowledge tests (such as ours), but they found no evidence that verbal tests produced more accurate self-estimates. In fact, they found that numerical tests produced the strongest (estimated r = .33 + .16 = .49). There is also a large body of research in industrial-organizational psychology that has investigated the role of question formats in self- and other-estimated abilities. Regrettably, ncertainty exists regarding what has been found, and what should be used in practice (DeNisi & Murphy, 2017). As such, this body of literature, though large, is unfortunately not immediately applicable to the interpretation of our results. Taken together then, it is unclear why we find larger correlations between measured and self-rated ability scores compared to other studies.

References

Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What we really know about our abilities and our knowledge. *Personality and Individual Differences*, *33*(4), 587–605. https://doi.org/10.1016/S0191-8869(01)00174-X

Alexander, S. (2013, April 12). *Lizardman's Constant Is 4%*. Slate Star Codex. https://slatestarcodex.com/2013/04/12/noisy-poll-results-and-reptilian-muslim-climatologists-from-mars/

Chalmers, P., Pritikin, J., Robitzsch, A., Zoltak, M., Kim, K., Falk, C. F., Meade, A., Schneider, L., King, D., Liu, C.-W., & Oguzhan, O. (2020). *mirt: Multidimensional Item Response Theory* (1.32.1) [Computer software]. https://CRAN.R-project.org/package=mirt

Dalliard, M. (2017, July 1). Measurement Error, Regression to the Mean, and Group Differences. *Human Varieties*. https://humanvarieties.org/2017/07/01/measurement-error-regression-to-the-mean-and-group-differences/

DeMars, C. (2010). *Item response theory*. Oxford University Press.

DeNisi, A. S., & Murphy, K. R. (2017). Performance appraisal and performance management: 100 years of progress? *The Journal of Applied Psychology*, *102*(3), 421–433. https://doi.org/10.1037/apl0000085

Freund, P. A., & Kasten, N. (2012). How smart do you think you are? A meta-analysis on the validity of self-estimates of cognitive ability. *Psychological Bulletin*, *138*(2), 296–321. https://doi.org/10.1037/a0026556

Gignac, G. E., & Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence*, *80*, 101449. https://doi.org/10.1016/j.intell.2020.101449

Harrell, F. E. (2019). *rms: Regression Modeling Strategies* (5.1-3.1) [Computer software]. https://CRAN.R-project.org/package=rms

Jensen, A. R. (1998). *The g factor: The science of mental ability*. Praeger.

Kirkegaard, E. O. W. (2021). Are there Complex Assortative Mating Patterns for Humans? Analysis of 340 Spanish Couples. *Mankind Quarterly*, *61*(3), 578–598. https://doi.org/10.46469/mq.2021.61.3.12

Krueger, J., & Mueller, R. A. (2002). Unskilled, unaware, or both? The better-than-average heuristic and statistical regression predict errors in estimates of own performance. *Journal of Personality and Social Psychology*, *82*(2), 180–188. https://doi.org/10.1037/0022-3514.82.2.180

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/10.1037//0022-3514.77.6.1121

Schimmack, U. (2020, September 13). The Dunning-Kruger Effect Explained. *Replicability-Index*. https://replicationindex.com/2020/09/13/the-dunning-kruger-effect-explained/

Shores, E. A., & Carstairs, J. R. (2000). The Macquarie University Neuropsychological Normative Study (MUNNS): Australian Norms for the WAIS-R and WMS-R. *Australian Psychologist*, *35*(1), 41–59. https://doi.org/10.1080/00050060008257467

Snow, W. G., Tierney, M. C., Zorzitto, M. L., Fisher, R. H., & Reid, D. W. (1989). WAIS-R test-retest reliability in a normal elderly sample. *Journal of Clinical and Experimental Neuropsychology*, *11*(4), 423–428. https://doi.org/10.1080/01688638908400903

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