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, indeed a simple statistically necessary empirical pattern, and the other an unsupported theory that purports to explain this pattern. Recently, (Gignac & Zajenkowski, 2020) suggested two ways to operationalize and test the theory. We carried out a replication of their study using archival data from a larger dataset. We used two measures of self-estimated ability: estimated sumscore (correct responses), and estimated own-centile. We find no evidence of nonlinearity for either. We find evidence of heteroscedasticity for self-centile estimates, but not raw score estimates. Overall, the evidence was mostly inconsistent with Dunning-Kruger theory.

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

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.
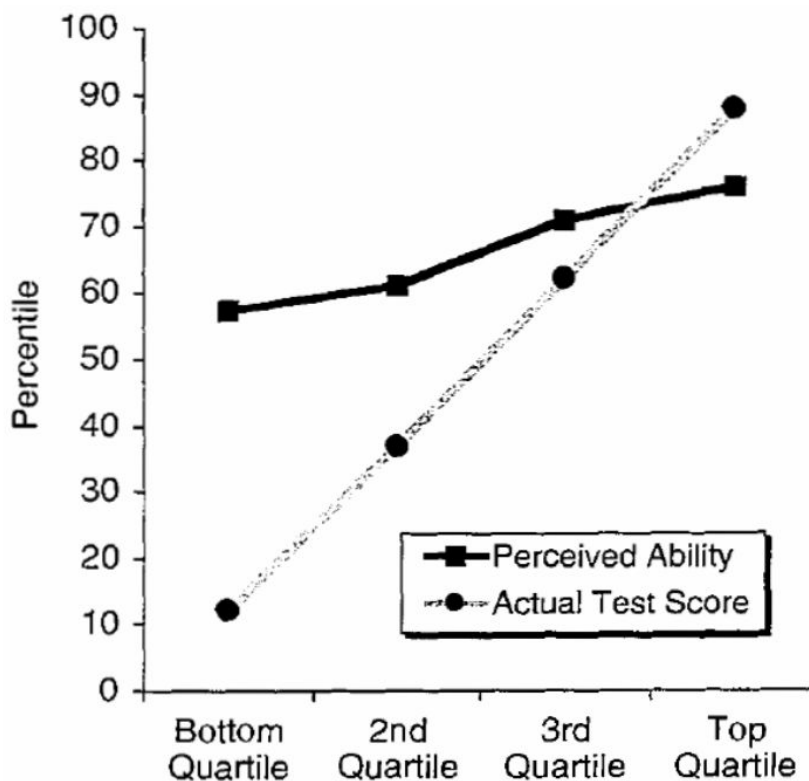

Figure X. Typical Dunning-Kruger pattern. Reproduced from (Kruger & Dunning, 1999).

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. Statistical criticism of the findings 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). Second, there is a general tendency to overestimate own performance in general (so called illusory superiority effect, though it may be limited to Europeans). When these two facts are combined, they yield the familiar Dunning-Kruger pattern, shown in Figure X.
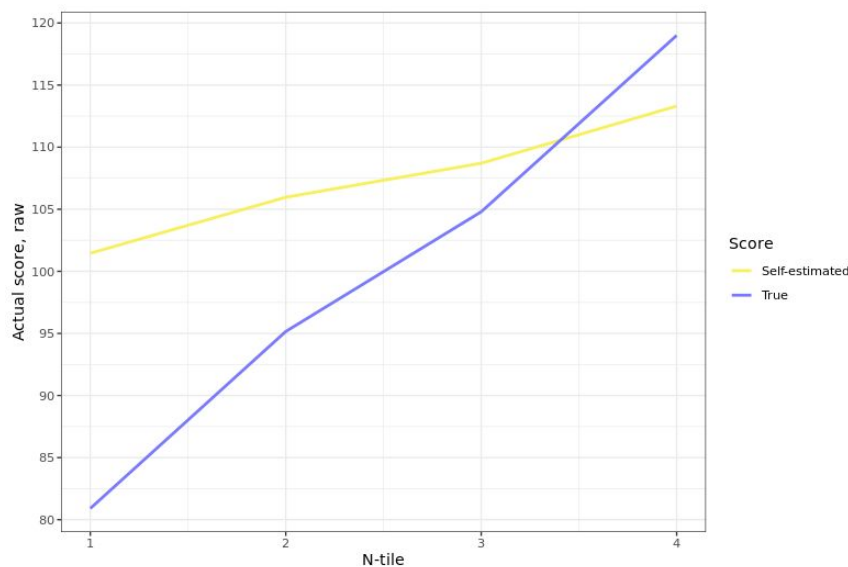

Figure X. An example of the Dunning-Kruger pattern. Based on simulated data from
http://emilkirkegaard.dk/understanding statistics/?app=Dunning Kruger.

As above, we see that people of below average ability nonetheless believe themselves to be above average. However, those high in ability tend to underestimate 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. However, recently, (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 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. 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 greater residual variance in the below average ability region. A second derived prediction is that the association between self-estimated and

real ability should diverge from the overall trend, with a weaker or zero 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. The purpose of this paper was to replicate the findings of (Gignac & Zajenkowski, 2020) in a new sample.

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). 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. 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?

Surprisingly, the estimates correlated only at .61, 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.
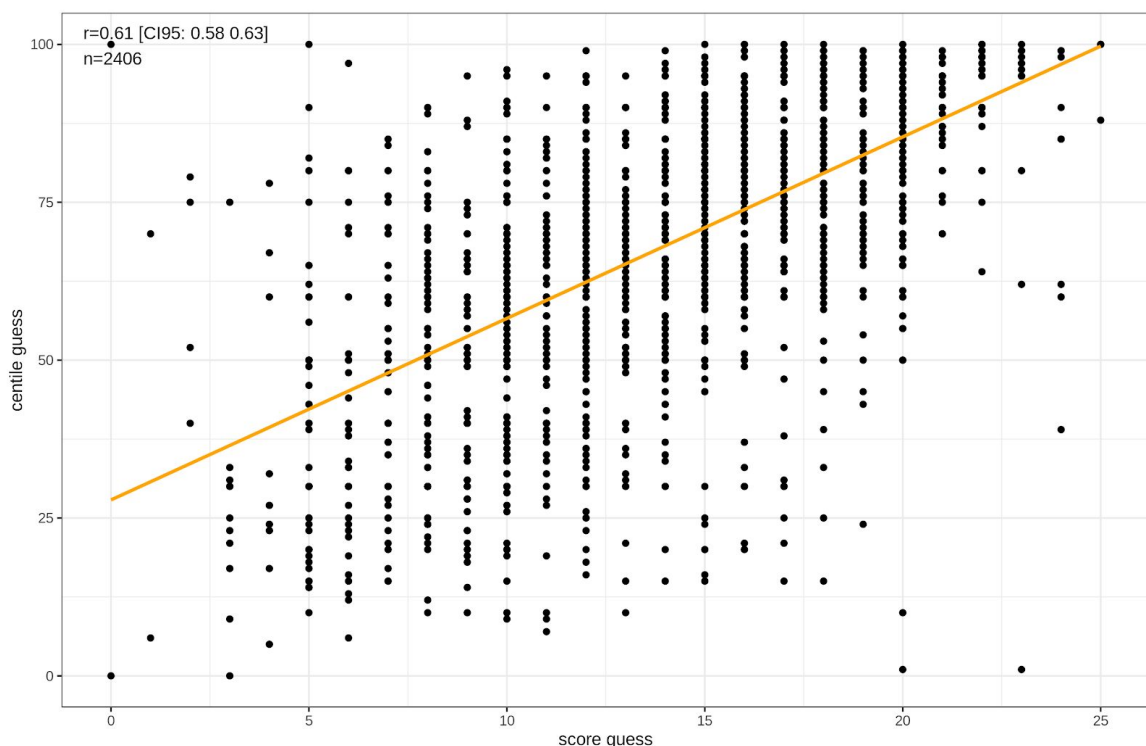

Figure X. Scatterplot of relationship between two different self-estimated ability variables.

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 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). Figure X shows the distribution of scores by scoring method. Their correlation was .95. The estimated reliability was .73 for the actual data, and .73 with an assumed perfect normal distribution.[1] These values are probably underestimates of the 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 pasted 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.
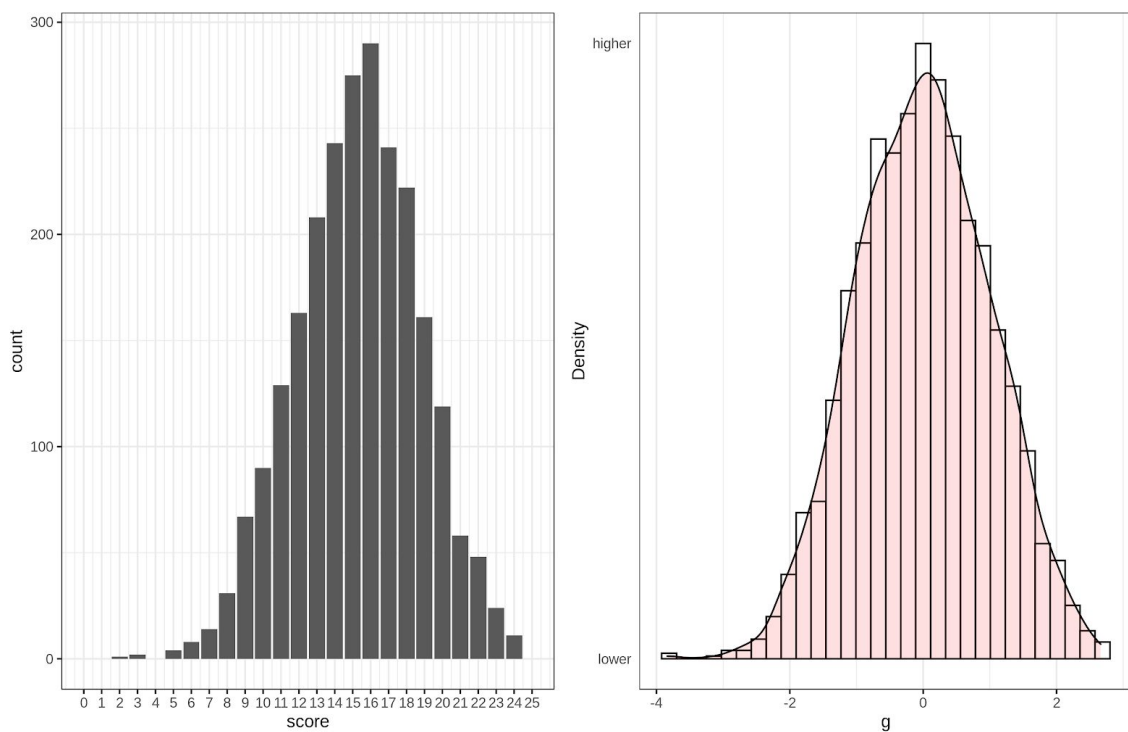

Figure X. Distributions of cognitive ability by scoring method. Left panel shows sum scores, and the right panel, item response theory standard scores (density curve overlaid).

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.

---

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

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.
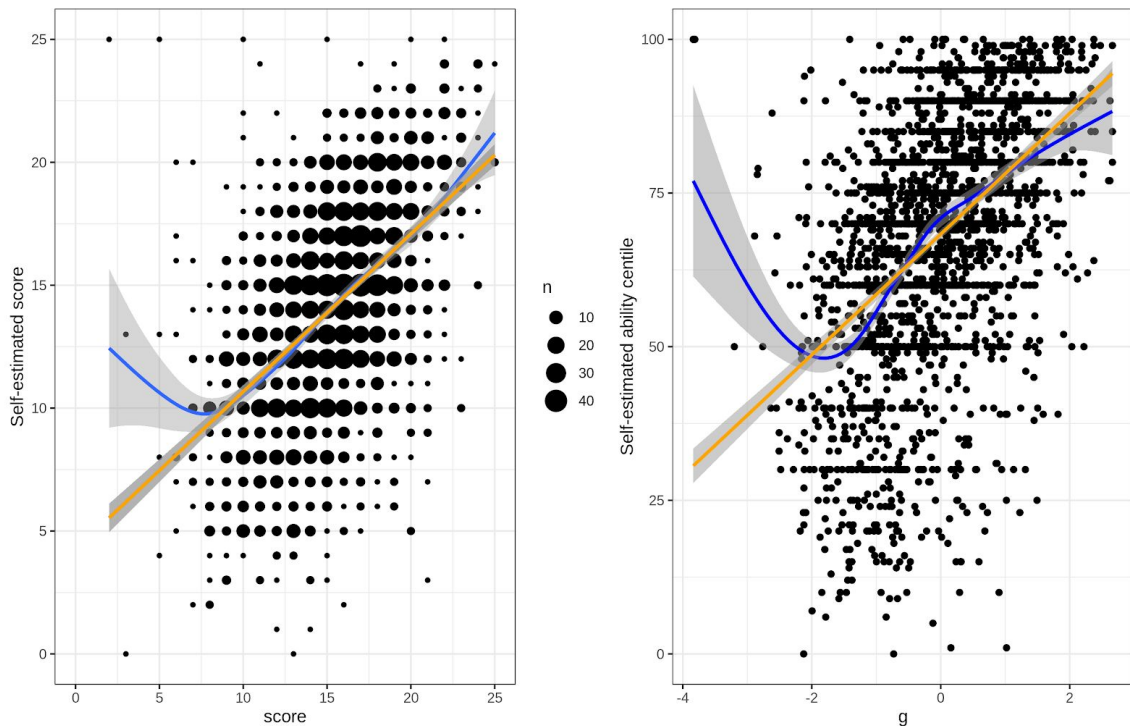

Figure Xa-b. Scatterplots showing linear (orange) and LOESS (blue; locally estimated scatterplot smoothing) fits for self-estimated ability and measured ability. Left plot: sumscore and self-estimated sumscore. Right plot: item response theory g score and self-estimated ability centile.

The correlations are relatively strong: r's .52 and .48, for score x self-estimated score, and g x 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).
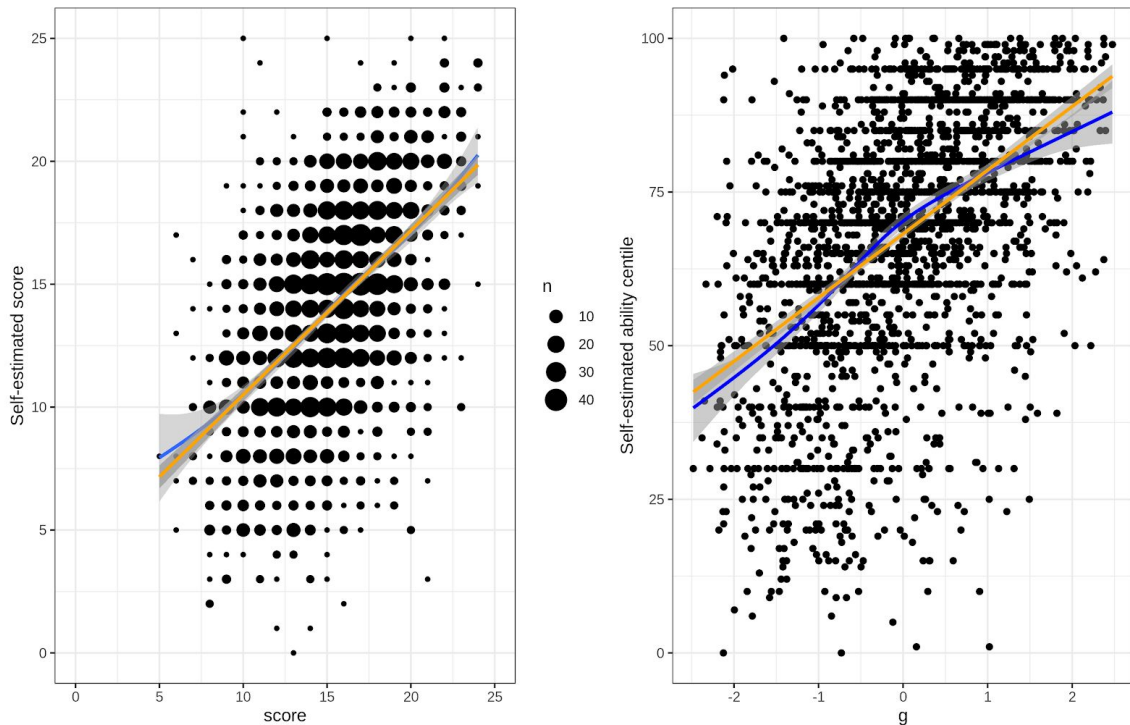
Figure Xa-b. Scatterplots showing linear (orange) and LOESS (blue; locally estimated scatterplot smoothing) fits for self-estimated ability and measured ability. Left plot: sumscore and self-estimated sumscore. Right plot: item response theory g score and self-estimated ability centile. Outliers beyond 2.5 z removed.

We see now in the plots that the association is now near-perfectly linear. To be fair, testing for nonlinearity using a model comparison (natural spline model vs. linear model) 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%).

Turning to the question of heteroscedasticity, we employed the same method as in (Kirkegaard, in press). In this approach, the model of interest is fit, then the residuals are standardized, and then converted to positive (absolute) values. Linear and nonlinear models are then fit to the residuals to look for evidence of heteroscedasticity. In the simulation study carried out by (Kirkegaard, in press), 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.
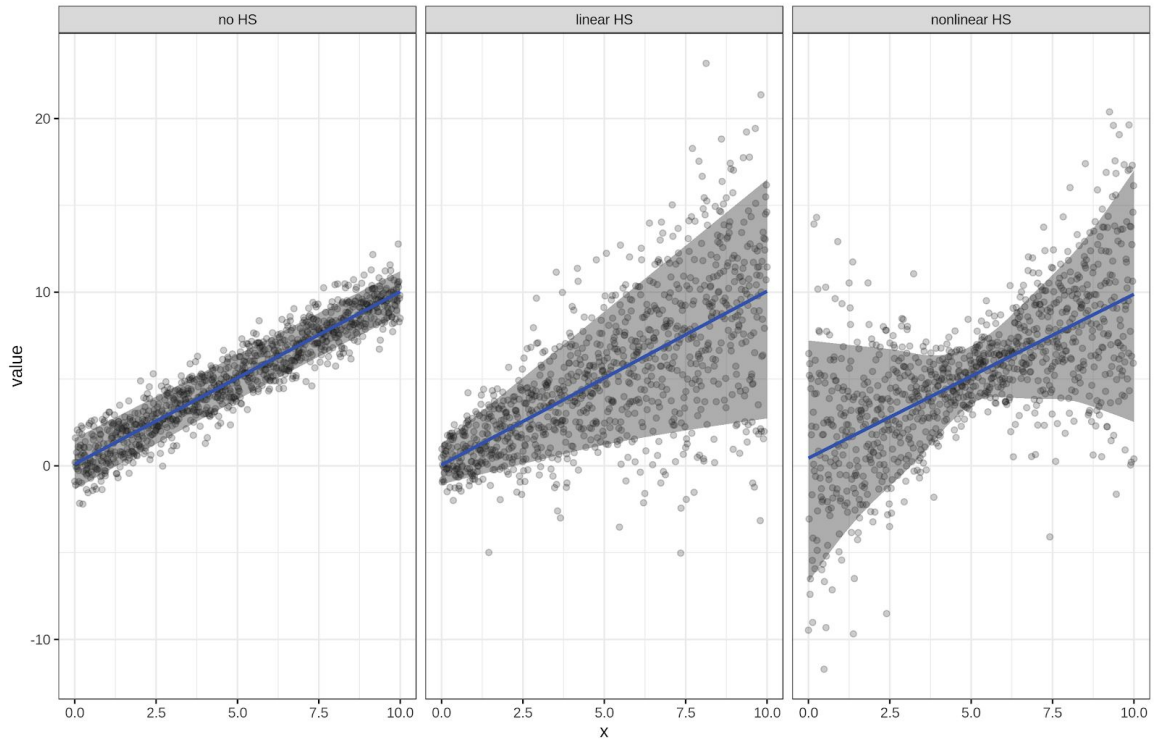
Figure X. Types of heteroscedasticity. Left: no heteroscedasticity (homoscedasticity), 2) monotonic linear increasing heteroscedasticity, and 3) nonmonotonic nonlinear heteroscedasticity.

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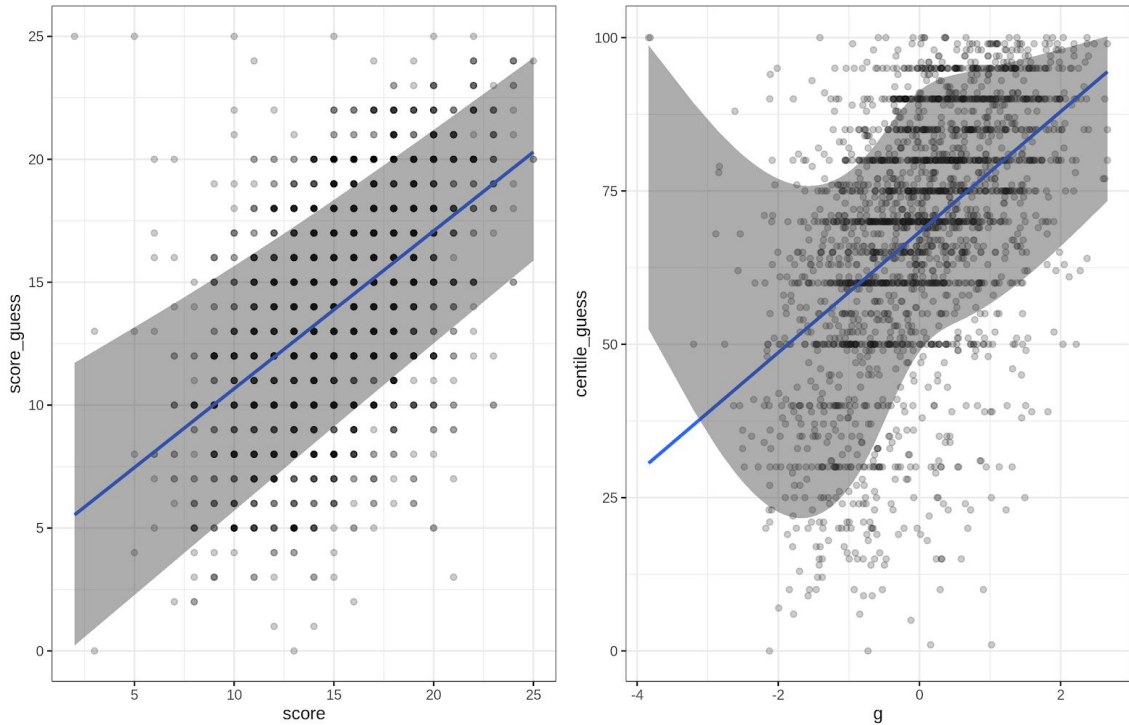
Figure X. Estimated 10th and 90th centiles of ability variables for the complete dataset.

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 seen in the prior section. Thus, we reran the tests on the reduced dataset. Figure X shows the results.
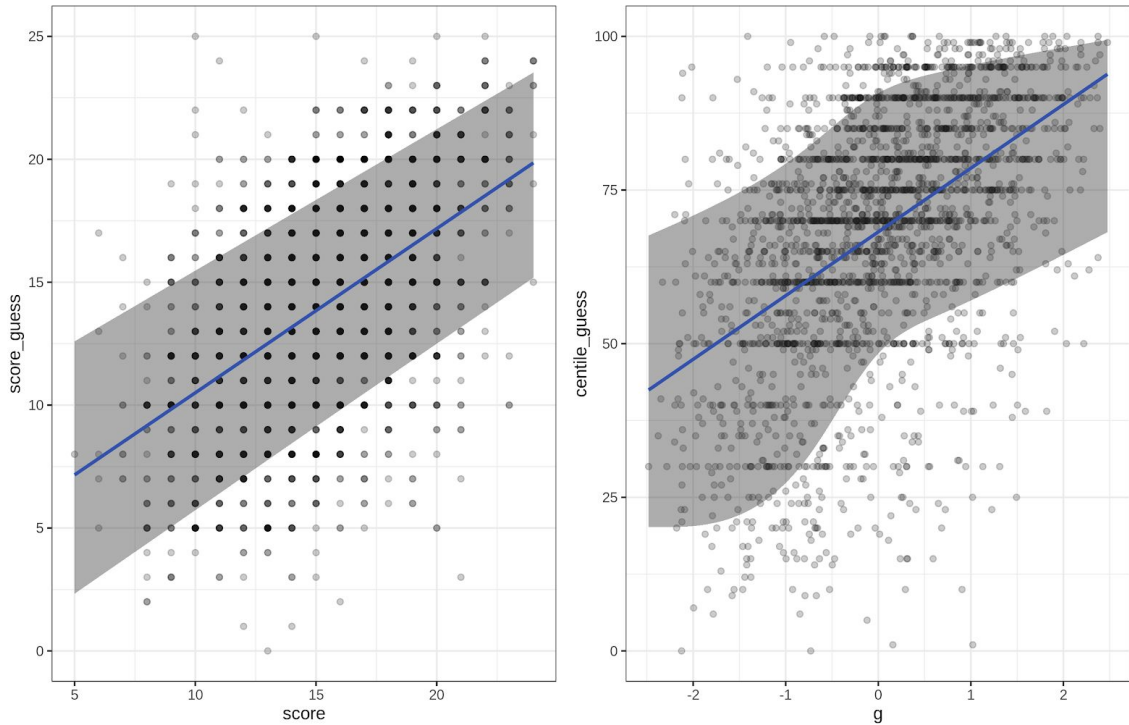
Figure X. Estimated 10th and 90th centiles of ability variables for the dataset with outliers removed.

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).

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	Sumscore1	Sumscore2	Sumscore combined	g 1	g 2	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 effect across two methods. First, our replication study was about 2.5 times larger than the study by (Gignac & Zajenkowski, 2020, n=929). For this reason, we have more statistical precision and our study carries more weight.

Second, since we had 2 self-estimation variables, we were able to conduct tests for both throughout the study. Because these two variables concern different aspects of self-estimation, they may be theoretically important. In our case, we find essentially no nonlinearity in the relationship to objectively measured ability once we account for a few left-tail outliers (Figure X). However, when we test for heteroscedasticity, we find evidence for the theory-predicted values only for the estimated centiles, not raw scores. The fact that centiles require the subject to also estimate the distribution of others' scores in order to estimate oneself may suggest the reason for this finding. Estimating one's own sumscore is merely a matter of summing the estimated probability for each item. Many people keep rough counts when taking tests for fun, so this mechanism is plausible. Here it should also be said that our online sample was probably somewhat above average, thus the centiles that we estimated are probably too low as compared to the general population. It is thus less surprising that our subjects were somewhat less precise on this task. Still, we find heteroscedasticity in the theory-predicted region, so this can be seen as partial validation of the Dunning-Kruger theory. In our opinion, the fact that below average subjects were somewhat less able to estimate centiles but not sumscores, is not something that Dunning-Kruger theory would have been able to predict beforehand. It does not seem like a failure of meta-cognitive theory, but rather just a general lower ability to estimate the likely score distribution and estimate own centile compared to this. In fact, Dunning and Kruger have themselves used variables that used

absolute score estimates to support their theory (Kruger & Dunning, 1999). Thus, overall, we interpret our evidence as being mostly inconsistent with Dunning-Kruger theory.

Third, we find 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 .47 to .54, depending on which combination of variables is used (cf. Table X). 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. Thus, we are not sure why we find a substantially higher correlation. Their moderator 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. However, uncertainty 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.

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-clim

atologists-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

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

Kirkegaard, E. O. W. (in press). Are there complex assortative mating patterns for

    humans? Analysis of 340 Spanish couples. *Mankind Quarterly*.

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