Comment on Gignac and Zajenkowski, "The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data"

Intelligence, 2023, 101732, ISSN 0160-2896, https://doi.org/10.1016/j.intell.2023.101732

Avram Hiller
ahiller@pdx.edu

Abstract: Gignac and Zajenkowski (2020) find that "the degree to which people mispredicted their objectively measured intelligence was equal across the whole spectrum of objectively measured intelligence". This Comment shows that Gignac and Zajenkowski's (2020) finding of homoscedasticity is likely the result of a recoding choice by the experimenters and does not in fact indicate that the Dunning-Kruger Effect is a mere statistical artifact. Specifically, Gignac and Zajenkowski (2020) recoded test subjects' responses to a question regarding self-assessed comparative IQ onto a linear IQ scale when a normal IQ scale would likely have been more appropriate. More generally, researchers studying self-assessed intelligence should be aware of potential measurement problems that may arise when transforming an ordinal scale onto an interval scale.

Keywords: Dunning-Kruger Effect; Mere Statistical Explanations; Self-Assessed Intelligence; Measurement Theory; Levels of Measurement

The Dunning-Kruger Effect (hereafter DKE) is the purported phenomenon that low-skill individuals in a domain lack higher-order knowledge that they are low-skilled (Kruger and Dunning 1999). However, a number of scholars (Krueger and Mulller 2002; Nuhfer 2016) have questioned whether the statistical evidence used actually confirms the existence of a DKE. Specifically, the evidence Kruger and Dunning (1999) and others have given for the DKE might just be explained as being a statistical artifact.

Gilles Gignac and Marcin Zajenkowski (2020, hereafter G&Z) insightfully employ two tests for statistical artifacthood: a Glejser test for homoscedasticity and a linearity test. G&Z conducted a study to show that under these tests, the DKE is indeed a mere statistical artifact. G&Z found, rather remarkably, that in a sample of 929 subjects, "the Glejser test correlation was near zero ($r = -0.05$), suggesting that the degree to which people mispredicted their objectively measured intelligence was equal across the whole spectrum of objectively measured intelligence" (6). In other words, subjects at the high end of the objective IQ spectrum do not predict their IQ

scores any better than those at the low end. Because the absolute residuals were normally distributed across the IQ spectrum, showing no heteroscedasticity, there is no evidence for a DKE. Furthermore, G&Z found a linear and not quadratic relationship between objective IQ and self-assessed IQ [*SAIQ*] scores, giving further reason to think that there is no evidence for a DKE.

However, G&Z's reported findings may be better explained not by the absence of the DKE but by their choice of how to recode subject responses to a question regarding SAIQ. Here is G&Z's description of their procedure (5):

> Participants' SAIQ was indexed with the marked column counting from the first to the left; thus, the scores ranged from 1 to 25. Prior to providing a response to the scale, the following instruction was presented:
>
> "People differ with respect to their intelligence and can have a low, average or high level. Using the following scale, please indicate where you can be placed compared to other people. Please mark an X in the appropriate box corresponding to your level of intelligence."
>
> In order to place the 25-point scale SAIQ scores onto a scale more comparable to a conventional IQ score (i.e., M = 100; SD = 15), we transformed the scores such that values of 1, 2, 3, 4, 5… 21, 22, 23, 24, 25 were recoded to 40, 45, 50, 55, 60… 140, 145, 150, 155, 160.

Is this a proper way to recode subjects' responses?

There is reason to believe that it is not. Given that subjects were asked to compare their IQs with others, it seems that it would be better to recode it not to a linear scale from 40 to 160 but onto the *actual* IQ scale, which I will here presume to be a normal distribution ($\mu = 100$; $\delta = 15$). For instance, what does a respondent who answers 20 (out of 25 on G&Z's scale) likely have in mind as their SAIQ? G&Z recode a 20 response as a 135 IQ, which is at the 99th percentile of IQ. But this seems like it would misrepresent what the subject has in mind. While we cannot expect respondents to know the normal IQ curve, more likely a person who responds with a 20 out of 25 on a comparative scale would have in mind that they are at about the 80th percentile, which is approximately an IQ of 112, rather than 135.

G&Z find that those at the high end of the IQ spectrum have a (surprisingly) large error rate. It should first be noted that unlike most DKE studies, G&Z found that on average, those at the highest end of the skill spectrum *over*estimate their IQs. More typically, as in Kruger and Dunning (1999), studies find that those at the highest end on average *under*estimate their IQs (likely in part because of regression to the mean). This discrepancy with previous studies provides

evidence that G&Z have coded the self-assessments of subjects in the highest end *considerably* above their actual intended self-assessments.

At the lower end of the skill spectrum, all studies agree that individuals overestimate their abilities. For a subject who responds 10 (out of 25) to the SAIQ question, G&Z would have coded that individual as having an SAIQ of 85, which is in fact at the 16th percentile. But more likely, the individual believes that their IQ is about at the 40th percentile, and thus their SAIQ would be better coded as a 96. Now, if this individual turns out to be objectively at, say, the 15th percentile, which is an IQ of 84.5, G&Z would have coded them as having a slight overestimated IQ of .5 points, but by my suggested coding, it would be an error of 11.5 points.

Thus, whereas G&Z's coding minimizes individuals' overestimates for those at the low end of the IQ spectrum, it greatly *increases* overestimates of those in the upper half (relative to my suggested recoding). If a person responded with a 23 (out of 25), seemingly indicative of being near the 90th percentile, which is an SAIQ of 119, G&Z would have coded the person as having an SAIQ of 150, which is in fact at the 99.96th percentile. Now, if this person's subjective IQ was in fact at the 85th percentile, which is an IQ of 115.5, a more proper coding would have it as an error of 3.5 points, but G&Z's coding would have it as an error of 34.5 points. It is thus highly likely that the combination of minimizing the overestimates of those at the low end of the spectrum and increasing the errors of those at the high end is the main reason why G&Z do not find that those at the low end have a higher absolute error rate than those at the high end.

Table 1 is a chart showing, for each of the 25 responses (Column 1), G&Z's SAIQ recoding (Column 2); the IQ percentile at that IQ (Column 3); the raw percentile of the subject response (Column 4);[1] my suggestion (Column 5) for how the response should have been coded based upon a normal distribution of IQ ($\mu = 100$; $\delta = 15$); and the difference between G&Z's coding and my suggested coding (Column 6).

1. Raw Response	2. G&Z Coded SAIQ	3. Percentile at G&Z IQ	4. Raw Percentile	5. IQ at Raw Percentile	6. SAIQ Difference
1	40	0.003	2	69.2	(29.2)
2	45	0.01	6	76.7	(31.7)
3	50	0.04	10	80.8	(30.8)
4	55	0.13	14	83.8	(28.8)

[1] Note that because the initial question scale is 1 to 25 rather than 0 to 25, and G&Z's 100 IQ midpoint is 13 (rather than 12.5), I have coded raw score responses *n* into percentiles (Column 4) by the formula IQ percentile = 4n – 2, to maintain the 50th percentile midpoint at n=13.

5	60	0.4	18	86.3	(26.3)
6	65	1.0	22	88.4	(23.4)
7	70	2.3	26	90.3	(20.3)
8	75	4.8	30	92.1	(17.1)
9	80	9.1	34	93.8	(13.8)
10	85	15.9	38	95.4	(10.4)
11	90	25.2	42	97.0	(7.0)
12	95	36.9	46	98.5	(3.5)
13	100	50.0	50	100.0	0.0
14	105	63.1	54	101.5	3.5
15	110	74.8	58	103.0	7.0
16	115	84.1	62	104.6	10.4
17	120	90.9	66	106.2	13.8
18	125	95.2	70	107.9	17.1
19	130	97.7	74	109.7	20.3
20	135	99.0	78	111.6	23.4
21	140	99.6	82	113.7	26.3
22	145	99.87	86	116.2	28.8
23	150	99.96	90	119.2	30.8
24	155	99.99	94	123.3	31.7
25	160	99.997	98	130.8	29.2

Table 1. Chart showing differences between G&Z's linear recoding and suggested recoding of SAIQ raw score.

Figure 1 is a graphical depiction of the data in Columns 2 and 5.

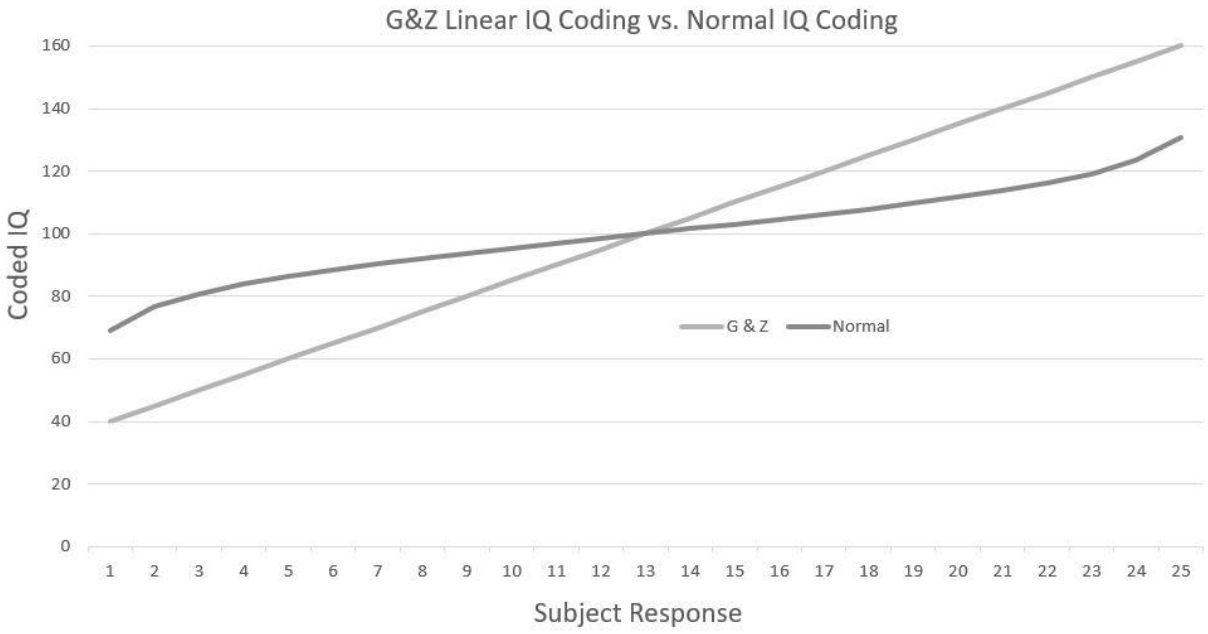

Figure 1. Graph depicting G&Z's linear coding and my suggested coding.

In sum, G&Z's raw data most likely do not show what they claim the data to show. My suggestion would be to recode G&Z's raw data in accord with the normal distribution as in Table 1, Column 5, and rerun the Glejser test and linearity test on it. Alternatively, G&Z or others could take raw data from other published tests of the DKE and run those two tests. It is worth noting that the recoding issue discussed here does not undermine G&Z's main *theoretical* point, which is that *if* residuals are found to be homoscedastic, or *if* the relation between objective IQ and SAIQ is linear rather than quadratic, then there would be no evidence that there is a DKE (though also see Gignac 2022, which questions whether the Glejser test is a valid test for this purpose).

I should note that Dunkel et al. (2023) have attempted to replicate G&Z (2020). Here is their method for determining self-assessed IQ (2023, §2.2.1):

> The *SAI* measure was comprised of two items. First, participants responded to the question "Compared to other people your age, how intelligent are you?". Responses were coded using a six-point Likert scale (1 = moderately below average; 2 = slightly below average; 3 = about average; 4 = slightly above average; 5 = moderately above average; 6 = extremely above average). Second, participants were later asked "How intelligent are you?" with responses indicated on a four point Likert scale (1 = very intelligent; 2 = moderately intelligent; 3 = slightly intelligent; 4 = not at all intelligent).
>
> A multi-step process was followed to create the *SAI* measure. First, item 2 ("How intelligent are you?") was reverse coded to match the coding direction of item 1. Second, both items were standardized (i.e., *z*-transformed) and then

> summed. Third, to produce estimated IQ scores, the total scores were then standardized once again, with a mean of 100 and a standard deviation of 15.

However, the process of summing these Likert questions and transforming them onto an SAIQ scale faces similar concerns as G&Z's original recoding. Dunkel et al. (2023) twice z-transformed the raw scores, but that linear transformation from their test questions onto SAIQ is unlike my suggested non-linear recoding. Furthermore, given that *SAIQ* itself has never been shown to have the same distribution as IQ (i.e., normal, with M = 100 and SD = 15), Dunkel et al.'s choice to recode SAIQ so that subjects' responses *as a whole* map onto the IQ curve (as opposed to my suggested recoding of *individual* responses onto the curve) is likely to introduce other discrepancies.[2]

I'd like, in conclusion, to make several further reflections on these issue. First, G&Z's method for determining self-assessed ability differs from standard methodology in studying the DKE. Most studies of the DKE, both those defending its existence (Kruger and Dunning 1999, Ehrlinger 2008, Jansen 2021) and those casting doubt upon it (Krueger and Mueller 2003, Burson 2006), simply assess subjective *percentile* (as expressed *directly* by respondents) against objective percentile. Additionally, Sophie von Stumm (2014) writes: "In psychological research, SEI [self-estimated intelligence] is typically assessed by showing participants a graph of an IQ score distribution… Participants then place themselves along a bell curve and report the corresponding IQ score value". This method seems precisely geared to avoid the problem that I am pointing out, since it explicitly shows the respondents the normal curve.

Second, there is a longstanding issue in Measurement Theory regarding the consequences of mapping an ordinal scale onto an interval scale (see Stevens 1946, 1955; Kampen and Swyngedouw 2000; Liddell and Kruschke 2018). Ordinal scales are those which lack a pre-established metric. One illustrative example is the numerical pain scale used by physicians. The Wong-Baker Scale has patients rate pain from 0-10; the relevant issue is whether the steps on the scale represent equal intervals. This matters clinically as physicians need to determine proper dosages of pain medication. There is some evidence that it is *not* an interval scale (see Oliveira et al. 2014; also see Myles et al. 1999).

[2] In personal correspondence, Curtis Dunkel wrote to me: "It seems to me that the easiest and most valid method is to not transform the data at all. We thought we would simply follow up on the original G&Z manuscript, hence we tried to follow their steps." I agree; for this reason, in this Comment I focus primarily on G&Z and not on Dunkel et al. (2023).

Now, I am assuming here that objective IQ itself is a legitimate interval scale, but SAIQ is not the same as objective IQ, and the main point is that there is *no pre-established, legitimized interval scale for self-assessed intelligence* to unproblematically map subjects' responses onto. (Perhaps it might be said that the very term used by G&Z, "self-assessed IQ", is itself something of a misnomer: respondents are not self-assessing their IQ; rather, respondents are self-assessing their intelligence and experimenters are mapping their responses onto IQ.) Ordinary people likely don't have a cognitive construct of an SAIQ elicited by the G&Z's specific question that the mapping would represent. On the other hand, von Stumm's method *gives* the respondent an IQ curve to place themselves upon, and so it is immune to this concern. And methods that ask subjects to rate themselves as a percentile are also seemingly immune to this concern.

In a more recent paper, Gignac (2022) asked respondents: 'On a scale from 1 to 7, where 1 means very low and 7 means very high, how would you assess your overall financial knowledge?'. Gignac then writes: "Responses were captured with a 7-point scale, ranging from 1 = very low to 7 = very high." Then, Gignac conducted a linearity test of objective financial knowledge against subjective financial knowledge, as in Figure 2.

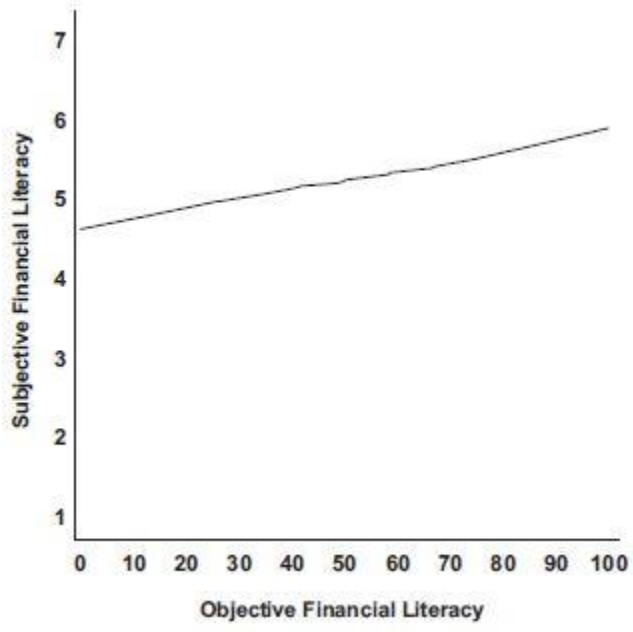

Figure 2. Response to a Likert question regarding self-assessed financial knowledge on y-axis graphed against objective financial literacy on x-axis. Source: Gignac (2022).

My concern is that it is not clear that the scale used for the y-axis is truly an interval scale, with equal increments between the numbers, and more evidence would be needed to show that it is. Given this, we should regard the linearity results of Gignac (2022) as not being fully established. Some (such as Knapp 1990) have argued that in some instances, it may be acceptable for practical purposes to treat an ordinal scale as an interval scale. However, when the very question at hand is whether the relationship between the two factors is linear (as in Gignac 2022), more should be done to show that the intervals on the y-axis are of equal units. Furthermore, there is no reason to think that there is a genuine underlying latent factor which could serve as an interval proxy for Gignac's (2022) ordinal scale,[3] and so there is insufficient reason to believe, prior to further empirical demonstration, that subjective financial literacy of respondents who are given a question on a seven-point scale from very low to very high will involve equal units.

All in all, given the considerations I have expressed here, I cannot with full assurance say that my *own* suggested recoding of G&Z's raw data onto the normal curve is unproblematic. It involves an empirical assumption about what is going on in test subjects' minds – that they intend their answers on G&Z's 1 to 25 scale to represent something akin to percentiles.[4] I have not directly tested this assumption, though the fact, noted above, that G&Z's data does not conform with previous studies provides some indirect evidence for it. So while I have raised concerns with G&Z's methodology, I have not demonstrated beyond any doubt that their recoding is erroneous. While I believe that G&Z's question is *likely* to be interpreted as being isomorphic to percentiles (albeit at ¼ of the range), my suggestion should be taken as providing what is most likely the best way to use G&Z's raw data, given how it was collected, but not as the best way to measure self-assessed intelligence in general.

Instead, my suggestion for future research on self-assessed intelligence, especially in relation to the Dunning-Kruger Effect, is for experimenters to either use von Stumm's (2014) method or simply to ask respondents to rate themselves in percentile terms, and then conduct

[3] See Kampen & Swyngedouw (2000). An example of this would be if experimenters asked subjects for how tall they think they are, from not at all tall to very tall, on a one to seven scale. Even though this is an ordinal scale, subjects' knowledge of their own height (in cm or inches) could demonstrate an underlying linearity. But even this is absent for self-assessed financial knowledge. Additionally, the x-axis scale of objective financial literacy was measured using short exams, and some care is also needed to ensure that there is a genuine internal trait of objective financial literacy that can properly be placed on an interval scale and is measured by the exams.

[4] Thanks to Gilles Gignac and Sophie von Stumm for raising this concern (personal correspondences).

statistical tests on the relation between subjective percentile and objective percentile.[5] (Another appropriate method also seen in the Dunning-Kruger literature – e.g. Feld 2017 – is to ask respondents to estimate their score on a test as opposed to their comparative level; see also Moore and Healy 2008 on the distinction between *misplacement* of one's ability relative to others and *misestimation* of one's score on a measure.) In sum, while this Comment focuses primarily on G&Z (2020), I believe it is important for researchers studying self-assessed intelligence more generally to be aware of the problems that may arise when an ordinal scale is used as, or recoded onto, an interval scale.

Acknowledgments: Thanks to Gilles Gignac for very helpful discussion, and to Sophie von Stumm and Curtis Dunkel for helpful correspondence. Thanks also to Liu-Qin Yang for discussion, and to the PDX PhiSciNow crew (Mark Bedau, Jay Odenbaugh, Anya Plutynski, and Wenqing Zhao) for detailed discussion of a draft of this Comment.

References

Burson, K.A., Larrick, R.P., & Klayman, J. (2006). Skilled or unskilled, but still unaware of it: how perceptions of difficulty drive miscalibration in relative comparisons. Journal of personality and social psychology, 90(1):60-77.

Dunkel, C.S., Nedelec, J., & van der Linden, D. (2023). Reevaluating the Dunning-Kruger effect: A response to and replication of Gignac and Zajenkowski (2020). *Intelligence*, *96*, 101717.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational behavior and human decision processes*, *105*(1), 98-121.

Feld, J., Sauermann, J., & De Grip, A. (2017). Estimating the relationship between skill and overconfidence. *Journal of behavioral and experimental economics*, *68*, 18-24.

Gignac, G.E. (2022). The association between objective and subjective financial literacy: Failure to observe the Dunning-Kruger effect. *Personality and Individual Differences*, 184, 111224.

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.

Jansen, R.A., Rafferty, A.N., & Griffiths, T.L. (2021). A rational model of the Dunning-Kruger effect supports insensitivity to evidence in low performers. *Nature Human Behaviour*, *5*(6), 756–763.

Kampen, J., & Swyngedouw, M. (2000). The ordinal controversy revisited. *Quality and quantity*, *34*(1):87-102.

Knapp, T.R. (1990). "Treating ordinal scales as interval scales: an attempt to resolve the controversy" *Nursing Res*. 39:121-123.

[5] I should note one concern for this method, which is that some subjects might not understand percentiles. If experimenters were to ask a comprehension question to exclude those who do not, it would then lead to a biased sample. For what it is worth, Burson et al. (2006) write "The use of the percentile scale was explained in detail."

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.

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.

Liddell, T.M., & Kruschke, J.K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong?. *Journal of Experimental Social Psychology*, *79*, 328-348.

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

Myles, P. S., Troedel, S., Boquest, M., & Reeves, M. (1999). The pain visual analog scale: is it linear or nonlinear? *Anesthesia and Analgesia*, *89*(6), 1517–1520. https://doi.org/10.1097/00000539-199912000-00038

Nuhfer, E., Cogan, C., Fleisher, S., Gaze, E., & Wirth, K. (2016). Random Number Simulations Reveal How Random Noise Affects the Measurements and Graphical Portrayals of Self-Assessed Competency. *Numeracy: Advancing Education in Quantitative Literacy*, *9*(1).

Oliveira, A.M., Batalha, L.M.C., Fernandes, A.M., Gonçalves, J.C., Viegas, R.G. (2014). Uma análise funcional da Wong-Baker Faces Pain Rating Scale: linearidade, discriminabilidade e amplitude. *Revista de Enfermagem Referência* 4(3):121-130. Note: English translation ("A functional analysis of the Wong-Baker Faces Pain Rating Scale: linearity, discriminability and amplitude") available at https://www.semanticscholar.org/paper/A-functional-analysis-of-the-Wong-Baker-Faces-Pain-Oliveira-Batalha/d735da3edb1ed0a249727e43cc2931f4226defdf.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677–680. https://doi.org/10.1126/science.103.2684.677.

Stevens, S.S. (1955). On the averaging of data. *Science*, 121(3135), 113–116. https://doi.org/10.1126/science.121.3135.113