Psychophysical deconstruction of the Dunning-Kruger effect

Robert D. McIntosh*, Elizabeth A. Fowler, Tianjiao Lyu, Sergio Della Sala

Human Cognitive Neuroscience, Psychology, University of Edinburgh, UK.

Author note: This is an archived copy of the original version of a paper, prior to peer review, submitted to the *Journal of Experimental Psychology: General*. Following peer review, the paper was improved, and the title was changed. The final version is published at http://dx.doi.org/10.1037/xge0000579.

An author's copy of the final paper, along with full data and analysis code, are archived at https://osf.io/8wjck/. Please cite the final paper, not the pre-print.

Preregistration and open data statement: The preregistered plans and materials for this study, with open data and analysis code are available at https://osf.io/ccgsz/

Acknowledgements: We are grateful to Alice Calder and Ink Pansuwan for assistance with data collection, and to Adam Moore and Steve Loughnan for comments on a draft of the manuscript.

*Corresponding author

Robert D McIntosh

Psychology, University of Edinburgh

7 George Square, Edinburgh, EH8 9JZ

Tel: +44 131 6503444

Fax: +44 131 6503461

Abstract

The Dunning-Kruger effect (DKE) is the finding that, across a wide range of tasks, poor performers greatly overestimate their ability, while top performers make more accurate self-assessments. The original account of the DKE involves the idea that metacognitive insight requires the same skills as task performance, so that unskilled people perform poorly *and* lack insight. However, typical global measures of self-assessment are prone to statistical and other biases that could explain the same pattern. We used psychophysical methods to examine metacognitive insight in simple movement and spatial memory tasks: pointing at a dot, or recalling its position after a short delay. We measured task skill in an initial block, and self-assessment in a second block, in which participants judged after every trial whether they had hit the target or not. Metacognitive calibration and sensitivity were indeed related to task skill, and partially mediated the DKE. In a second study, we again measured task skill in an initial block, but titrated task difficulty in the second block so that all participants performed the task with equivalent levels of success. Metacognitive measures were again related to skill, but the DKE pattern itself was eliminated. In a third study, we used statistical modelling to illuminate these findings, showing that differences in metacognitive calibration and sensitivity can contribute to the DKE, but are neither necessary nor sufficient for it. This general analysis explains and quantifies how metacognitive insight and other factors interact to determine this famous effect.

Introduction

"*The fool doth think he is wise, but the wise man knows himself to be a fool.*"

William Shakespeare, *As You Like It*, Act V, Scene I.

This quotation, like others expressing similar ideas, has the appeal of instant connection. We can readily bring to mind examples that fit the template; even if we temporarily overlook counter-examples of the diffident fool or the arrogant genius. In experimental psychology, the idea is encapsulated by the Dunning-Kruger effect (DKE), the finding that, across a wide range of tasks, the poorest performers greatly overestimate their own ability, whilst the top performers make more accurate self-assessments. This statement of inverse correlation might lack the poetry of Shakespeare, but it has gained an almost viral momentum in contemporary discourse, particularly in the wake of the 2016 US Presidential election. But, even as the DKE has been embraced by the wider public, there has been debate within the psychological literature. The empirical pattern is not in doubt; the basic fact of relatively more overestimation amongst the objectively poorest performers is robustly replicable across a wide range of cognitive and social tasks (see Dunning, 2011), and has recently been extended to the domain of political beliefs (Hall & Raimi, 2018). The debate is about the correct explanation for the effect, and in particular whether the DKE implies metacognitive differences between the skilled and the unskilled in a given domain.

The original account offered by Kruger and Dunning (1999) hinges on the premise that, for many tasks, accurate appraisal of one's own performance depends on the same skills required for accurate performance. For instance, to judge the grammaticality of a sentence correctly, one must have the grammatical knowledge needed to compose it. Kruger and Dunning argued that the lowest performers in a task suffer a 'dual burden': not only is their performance poor, but they have a corresponding metacognitive deficit that impedes the ability to distinguish accurate from inaccurate performance. Unable to discern their own errors, poor performers assume they make fewer errors than in fact they do, resulting in overestimation. Conversely, high performers, with better task skills, have more metacognitive insight, so make better-calibrated self-estimates. However, when estimating themselves relative to others, high performers may still fall prey to a 'false consensus' effect, mistakenly assuming that other people's abilities are more similar to their own than they really are. High performers thus tend to underestimate themselves when using relative scales, though they

show less consistent under- or over-estimation when making absolute estimates (Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008). Low performers tend to overestimate themselves on both relative and absolute scales. At its broadest, the DKE is characterised by greater overestimation of own performance in low performers relative to high performers, that is, by a negative correlation between task skill and estimation error.

Competing accounts of the DKE focus on other mechanisms that might induce this negative correlation. One mechanism often invoked is regression to the mean (Ackerman, Beier, & Bowen, 2002; Burson, Larrick, & Klayman, 2006; Feld, Sauermann, & de Grip, 2017; Krueger & Mueller, 2002; Nuhfer, Cogan, Fleisher, Gaze, & Wirth, 2016; Nuhfer, Fleisher, Cogan, Wirth, & Gaze, 2017). A common method for studying the DKE involves ranking people by task performance and examining the relationship with estimation error, as indexed by the subtraction of actual from estimated performance; a negative correlation is virtually guaranteed in this scenario; just as, for any two imperfectly correlated random variables, x correlates negatively with y-x. In concrete terms, assuming imperfect self-estimation, then any chance variations that increase errors will be inadequately tracked in self-estimation, pushing people down the ranks of performance and simultaneously biasing them towards over-estimation; and vice-versa for chance variations that reduce errors. This artefact is important to recognise, but relatively easy to eliminate, either by quantifying and controlling for unreliability in the measure of performance (Ehrlinger et al., 2008; Krueger & Mueller, 2002), or by using separate sub-sets of trials to index performance and to calculate estimation error (Burson et al., 2006; Feld et al., 2017; Klayman, Soll, González-Vallejo, & Barlas, 1999). If either of these steps is taken, the DKE is reduced in strength, but it is not eliminated, so further factors must also be at work.

One such factor may be another ‘regressive’ tendency, which can persist even when regression to the mean is controlled for. If participants have imperfect knowledge of their performance, their estimation errors need not be random, but may be systematically biased. For many abilities and tasks, at least those that are relatively easy, people tend to evaluate their relative standing optimistically, a bias known as the better-than-average effect (see Kruger, 1999). If self-estimates are biased toward a high percentile, then higher performers will *appear* to be better-calibrated just by happening to perform closer to that percentile (Krueger & Mueller, 2002). This idea can be generalised to more difficult tasks, in which people’s relative estimates may instead show a pessimistic, worse-than-average effect (Kruger, 1999). Burson and colleagues (2006) contrasted difficult and easy tasks, for instance

questions about years of Nobel prizes requiring precise (within 5 years) or approximate (within 30 years) answers. They broadly confirmed that increased difficulty reduced the average percentile self-estimate, so that poor performers now seemed better calibrated, reducing and even reversing the asymmetry of the classic DKE.

Burson and colleagues (2006) argued that a sufficient account of the DKE is given by uncertainty of self-estimation combined with a bias towards a task-specific default estimate. This account was concerned with *relative* estimates, but it has been pointed out that the DKE is not restricted to relative estimates. Poor performers also tend to overestimate themselves when giving absolute estimates (see Dunning, 2011). However, there is another regressive bias that could promote greater overestimation amongst poor performers, even on absolute scales. This is a form of contraction bias, most clearly expressed by the fact that a person who performs a task perfectly (or perfectly poorly) can misestimate in one direction only. A poor performer is prone to overestimate simply because they have more errors than successes about which to be mistaken, and vice-versa at the top end (Krajč & Ortmann, 2007). To uphold an account in terms of metacognitive differences between low and high performers, it is not enough just to show that the effect persists once regression-to-the-mean has been controlled for (e.g. Ehrlinger et al., 2008), or when absolute self-estimates are used (Dunning, 2011). It is necessary to positively demonstrate these metacognitive differences, and to show that they causally mediate the DKE.

The original report by Kruger and Dunning (1999, Study 4) did include a proposed measure of metacognition. After completing a ten-item logical reasoning task and giving global ratings of how they thought they had performed, participants returned to their test sheets and tried to identify, item-by-item, which questions they had answered correctly. 'Metacognitive skill', measured by summing the total number of accurate identifications, correlated strongly with task performance and with estimation error, and the latter relationship persisted even when the former was controlled for. However, Krueger and Mueller (2002) pointed out that this pattern could be also be explained by a general optimism bias. For instance, if all participants guessed that they had performed at 60% correct, with no metacognitive insight, they would mark six of the ten answers as correct at random. Every objectively correct item would have a .6 chance of being identified accurately, whilst every incorrect item would have only a .4 chance of being identified accurately, so the overall accuracy of chance identifications would increase linearly with objective performance. Krueger and Mueller (2002) went on to replicate the mediational role of this measure of

metacognition, but they showed that it disappeared for an adjusted score that took account of the confound with performance.

Burson and colleagues (2006) took a different approach to metacognition, distinguishing between metacognitive *calibration* and *sensitivity*. The typical measure of estimation error is concerned with calibration (divergence of estimated from actual performance), but is unreliable, because a person with no insight can appear to be well-calibrated by guessing a value that happens to be close to their true performance. They proposed an alternative measure, of metacognitive *sensitivity*, given by the degree to which estimated performance tracks actual performance across participants. They calculated this correlation separately for top- and bottom-half performers. The results were mixed but, across three studies, bottom-half performers tended to show weaker correlations (lower sensitivity). However, Burson and colleagues warned that this correlation-based measure was also vulnerable to distortions. For instance, if the range of performance was more compressed amongst lower-half participants, then the measure of metacognitive sensitivity would be artificially reduced. They concluded that there was tentative evidence for reduced metacognitive sensitivity amongst poor performers, but that this did not imply poorer metacognitive calibration, and did not explain the DKE.

The original hypothesis, that the DKE is due to differences in metacognitive insight linked to task ability, thus remains open to debate. Indeed, no study (to our knowledge) has had metacognitive measures that could furnish a fair test of this hypothesis. In establishing such measures, we pick up on Burson and colleagues' (2006) conceptual distinction between *sensitivity* and *calibration* as aspects of metacognitive insight. Sensitivity implies an ability to detect variations in one's performance, whilst calibration reflects the correspondence or divergence between one's subjective criterion for success, and the criterion used by external judges. If we combine this with item-by-item self-estimation, we can frame a classical psychophysical analysis. The metacognitive task is to discriminate successful items or trials (hits) from unsuccessful ones (misses), given a varying signal strength (size of response error on some task-relevant dimension). Over sufficient trials, we can fit a logistic function to the probability of hit reports across levels of response error. If the participant has no insight, or if the response errors do not span the subjective transition from hits to misses, the fit will fail. But if the fit is good, then the function will define a threshold (at which the probability of reporting a hit is .5), and a just noticeable difference (for a .25 change in probability of reporting a hit). The threshold will quantify the participant's subjective criterion for success

(calibration) and the just noticeable difference will quantify their sensitivity. These metacognitive measures are in principle independent from performance, providing an adequate basis for testing the relation between task performance and metacognition, and the role of metacognition in determining the DKE.

One reason that a psychophysical approach has not previously been applied to this phenomenon may be because the tasks typically studied are high-level knowledge-based or reasoning tasks, in which answers can be scored as right or wrong, but for which it might be hard to quantify the degree of error (e.g. in terms of conceptual distance from the correct answer). Moreover, test items may be complex, making it onerous to collect sufficient cognitive and metacognitive responses for a psychophysical approach. Usually, participants are asked to estimate their performance globally, not trial-by-trial. Kruger and Dunning's (1999) original study did include item-by-item reports for a logical reasoning task, but for ten items only. In the present study, we move away from higher-level intellectual skills, to the more tractable domains of movement and spatial memory: pointing directly at a dot on a screen, or recalling its position after a short delay. These tasks give continuous measures of response error, and permit large numbers of trials, with online (trial-by-trial) reports of perceived success or failure, sufficient for a psychophysical analysis of metacognitive insight at the level of the individual participant.

The tasks are simple, but they fulfil the core requirements for the DKE to emerge. First, they are within the competence of participants, yet not at floor or ceiling, so that between-participant variations in performance are obtained. Second, test-retest reliabilities are sufficiently high that these variations can be meaningfully linked to differences in skill. Third, the available feedback is sparse enough that participants have some uncertainty about their performance. Finally, it is plausible that metacognitive insight on these tasks might depend on the same core competencies as performance (at least as plausible as for higher-level intellectual skills). For instance, to point accurately to target dots, one must have a precise forward model of movement, to enable the rapid detection and correction of errors. Participants with more precise forward models will be more accurate in pointing, but also better equipped to evaluate their accuracy. In remembering dot positions, participants with a more precise spatial memory will be more accurate, but may also have more diagnostic variation in their sense of certainty on occasions that their representation is less precise.

In Experiment 1, we propose to assess the DKE in movement and memory, using online self-estimation across *hundreds* of trials. This will allow for a global analysis of estimation error, to determine whether the DKE generalises to these novel tasks. Crucially, it will also support a rigorous, theoretically-grounded, psychophysical analysis, to quantify metacognitive calibration and sensitivity in every participant. These innovations allow us to mount the first adequate test of the now-famous, original account of the DKE: that poorer performers have less metacognitive insight, and that this mediates the inverse correlation between task skill and estimation error.

Experiment 1: Methods

Our tasks were developed through extensive piloting. Our methods and predictions were then preregistered on the Open Science Framework https://osf.io/ccgsz/. We report how we determined our sample size, all data exclusions, all manipulations, and all measures.

Participants and power

A sample size of 84 would provide .80 power to detect a correlation of .30, our minimum expected effect size of interest, based on pilot data. A total of 101 healthy participants were recruited amongst students and alumni of the University of Edinburgh, mostly in their early 20's (min 18, 1st quartile 20, median 21, 3rd quartile 23, max 42 years), and mostly female (82 female, 19 male) and right-handed (94 right-handed, 7 left-handed, by self-report). After exclusions, the final sample had 80 participants for the Movement task, and 62 for the Memory task.[1]

Procedure

Each task, Movement and Memory, had a baseline block and a main block. The purpose of the baseline block was to provide experience of the task, and to obtain an independent measurement of performance. Task order (Movement or Memory first) was alternated between participants. In each task, participants sat in front of a touchscreen (340 x 270 mm, 1024 x 768 pixels, ~0.33 mm per pixel), under dim ambient lighting, with the preferred hand resting on a start button 350 mm from the screen, and ~150 mm in front of the body midline. The Memory task also used a wireless mouse, to the preferred side of the start button.

Movement task. On each trial, a white dot was shown on a black screen, with its centre positioned randomly within a 700 pixel virtual square around the screen centre. The dot size was medium (14 pixel radius) in the baseline block, and small or large (10 or 18 pixel radius) in the main block. Dot presentation was initiated by the participant depressing

[1]Data were lost for three participants in the Movement task, and two participants in the Memory task, due to computer errors at testing. Fifteen participants failed to complete the Movement task due to an excess of time-errors. Three further participants were excluded for the Movement task, and thirty-seven participants for the Memory task, due to an inability to fit a significant binomial logistic regression to online self-estimations.

the start button, and the dot disappeared as soon as the participant released the button to initiate a response. The response was a paced reaching movement to the position of the dot, aiming to synchronize arrival with an auditory tone (100 ms, 500 Hz, 450 ms after button release). If a touch was registered within 350 ms, or if no touch was registered within 500 ms, a time-error message ("TOO FAST" or "TOO SLOW") was shown, and the trial was recycled. This narrow time window was imposed to limit the scope to trade speed against accuracy, ensuring that differences in task skill would be measurable in terms of accuracy. The baseline block continued until 100 valid responses were recorded, or was aborted after 150 total trials.

If the baseline block was completed successfully, the participant progressed to the main block. Dot size (small or large) on each trial was selected pseudo-randomly, and the main block continued until 100 valid trials were recorded for each target size, or was aborted if more than 300 trials were performed in total. After each valid movement, a dialog box appeared with two buttons, the upper (green) button labelled "HIT" and the lower (red) button labelled "MISS". The participant had to press one of the two buttons to report whether they thought that they hit the target location or not, providing an online record of self-estimation.

Online self-estimation is our focus in the present study; but we also collected more standard global estimates, immediately before and after the main block. The prospective estimates were absolute, with one rating screen for each dot size (small and large), with the wording, "MOVEMENT TASK: YOU WILL HAVE HALF A SECOND TO REACH AND HIT A DOT OF THIS SIZE. OUT OF 100 ATTEMPTS, HOW MANY TIMES DO YOU THINK YOU WILL HIT THE DOT?". The experimenter emphasised that the question related only to movements that were on-time, and that a 'hit' was a touch at the place that the dot had been shown. The participant touched a horizontal scale (0-100) to make their estimate, and a line appeared at the touched location. The participant could revise their estimate by retouching, or press "submit" to record the response. The order of rating screens (small or large dot first) was alternated between participants. For the retrospective ratings, the first two screens were identical to the prospective ratings except that the wording was in the past tense. Two further retrospective screens then asked for relative (percentile) estimates for each dot size: "OUT OF 100 HEALTHY ADULTS DOING THIS TASK, HOW DO YOU THINK YOU WILL COMPARE, IF 0 IS WORST AND 100 IS BEST?".

Memory task. The Memory task followed a similar format except that the instruction was to memorize the position of the dot and then release the button. Once the button was released, a dynamic white-noise mask filled the screen for 1000 ms, after which the screen returned to black except for a white crosshair cursor (6 pixel radius) at the screen centre. The participant used the mouse to guide the cursor to the remembered position, clicking to confirm their response, with no time limit. The prospective and retrospective self-estimates were identical to those for the Movement task, except for the precise wording, for instance, "MEMORY TASK: YOU WILL HAVE TO REMEMBER AND CLICK THE POSITION OF A DOT OF THIS SIZE. OUT OF 100 ATTEMPTS, HOW MANY TIMES DO YOU THINK YOU WILL HIT THE DOT?".

Data treatment and dependent variables

The first ten valid trials of baseline blocks were discarded as practice. For every other valid trial, response error was expressed as the number of pixels deviation from the boundary of the dot, with responses within the boundary coded as negative and responses beyond the boundary as positive. Response error was then converted into a binary hit (1) or miss (0), defining a hit within a six pixel penumbra around the dot.[2]

The percentage hit rate in the baseline block is a measure of performance that is independent from the calculation of estimation errors, and provides our general index of task skill. Performance in the main block is used in the calculation of estimation error. There were four global measures of *self-estimated performance*. The online self-estimations provided an overall online estimated hit rate, and the global rating screens provided prospective and retrospective absolute estimates, and a retrospective relative estimate. To convert these into *estimation errors*, we subtracted the actual hit rate in the main block; or, for the relative estimate, the main block hit rate expressed as a percentile relative to other participants in the analysis. Positive estimation errors reflect overestimation and negative estimation errors underestimation.

[2]Pilot testing with fully visible dots found that this penumbra was needed for an effective alignment of the objective criterion with the subjective impression of the fingertip overlapping the dot; and, in the Memory task, six pixels was the radius of the response cursor.

We also modelled the psychophysical relationship between online self-estimation reports and response error. Figure 1 depicts the analysis for one participant. We fitted a binomial logistic regression to predict the self-estimation report (0 or 1 for miss or hit) from the response error. We then calculated the subjective error threshold (SET) for distinguishing a hit from a miss, as the response error for which the probability of reporting a hit was .5. Lower SETs represent more conservative criteria for success and higher SETs more liberal criteria (a SET of six would represent perfect calibration to the objective criterion for a hit). We calculated the just noticeable difference (JND), following convention, as half of the stimulus (response error) difference associated with a change in the probability of reporting a hit between .75 and .25 (i.e. the semi-interquartile difference). Lower JNDs reflect steeper psychophysical functions, representing higher sensitivities of self-estimation; higher JNDs reflect more shallow functions, representing lower sensitivities.

Participants were excluded at the binomial logistic regression stage if they had fewer than ten estimation responses available in either category (hit or miss), or if the regression did not find a significant effect of response error on self-estimated performance, as assessed by the Wald test for that predictor ($p > 0.05$). In these cases, the measures SET and JND could not be meaningfully estimated.

Our main inferential analyses use Spearman rank correlation coefficients, as this makes minimal distributional assumptions, allowing us to pre-specify our analyses fully, and to avoid data transformation and unnecessary exclusions. Kruger & Dunning's original (1999) report of the DKE divided participants into sub-groups using performance quartiles. However, where we wish to form sub-groups, we will use performance tertiles (low, middle, high). This allows more participants per subgroup, but should not alter the overall patterns observed.

Experiment 1: Results

Estimation errors and performance

The DKE is assessed via the relationship between performance and estimation errors. The expected pattern is a negative correlation, with poorer performers showing more overestimation than good performers. In the present study, we have two measures of performance (hit rate in the main block, and in the baseline block), and four measures of estimation error (online, prospective absolute, retrospective absolute, and retrospective relative). All eight pairings of performance and estimation error are plotted in Figure 2a for the Movement task, and in Figure 2b for the Memory task. Each panel shows the mean estimation error for each tertile of performance, and the correlation across all participants.

The correlations are all negative, but vary in strength. Negative correlations are stronger if the measure of performance is taken from the main block (upper rows in Figures 2a and 2b). This is expected, because this measure of performance is also used in the calculation of estimation error, so these correlations are prone to regression to the mean. To remove this artefact, we should index performance by hit rate in the baseline block. Baseline performance is sufficiently related to that in the main block (ρ = .58 and .79 for Movement and Memory tasks respectively), that we can meaningfully treat baseline performance as an index of task skill. When estimation errors are plotted as a function of baseline performance, the pattern of negative correlation persists, albeit at a reduced level (lower rows in Figures 2a and 2b).

Negative correlations are generally stronger when self-estimation is relative, presumably because a relative estimate is affected not only by uncertainty over one's own performance, but also by uncertainty over other people's. To the extent that the DKE is driven by uncertainty, it will be inflated for relative estimates. Negative correlations persist when self-estimates are absolute, though these are generally more modest, and the tendency to under- or overestimation in top performers is less consistent, as previously noted (Dunning, 2011; Ehrlinger et al., 2008). Most importantly for present pruposes, negative correlations are obtained for online self-estimation (lower left panels of Figures 2a and 2b; ρ = -.30 and -.58 for Movement and Memory tasks respectively).

Online self-estimation

	Main block hit rate	Estimated hit rate	Online EE	SET	JND
Main block hit rate	-	.09	-.60	-.49	-.56
Estimated hit rate	.20	-	.71	.74	-.25
Online EE	-.74	.45	-	.95	.17
SET	-.38	.70	.86	-	.27
JND	-.75	-.15	.57	.45	-

Table 1. *Experiment 1. Spearman's ρ for correlations amongst performance and online self-estimation measures in the main block for the Movement task (n=80) (unshaded upper cells), and the Memory task (n=62) (shaded lower cells). The significance threshold would be ≥ .22 for the Movement task and ≥ .26 for the Memory task (two-tailed, uncorrected, alpha = .05).*

We now focus on online self-estimation. Table 1 shows the inter-correlations amongst key measures from the main block. In both tasks, the relationship between hit rate and online estimated hit rate, was non-significant (Movement task, $\rho = .09$, $n = 80$, $p = .45$; Memory task, $\rho = .20$, $n = 62$, $p = .15$). At face-value, participants seem to have no insight into their own performance. However, lack of insight at an individual level cannot be inferred from this result, because global estimation error conflates possible influences of metacognitive sensitivity and calibration (and other task-induced biases: Burson et al., 2006; Krueger & Mueller, 2002). Our online self-estimation method allows us to disentangle these aspects of metacognition, via the measures SET and JND. The fact that these measures could be meaningfully extracted for the majority of participants actually demonstrates that they did have significant insight into their performance.

In both tasks, participants with higher hit rates showed sharper sensitivity to response error (lower JND), and a more conservative SET criterion for claiming a hit. Table 1 further

shows strong positive relationships between online estimation error and our metacognitive measures, especially SET. This pattern, in which metacognitive sensitivity and calibration are associated both with task performance *and* with estimation errors, makes it possible that they could partially or wholly account for the DKE.

We assess the DKE with respect to baseline performance, to eliminate regression to the mean. Figure 3 shows how baseline performance relates to our online self-estimation measures. The top row shows the full scattergrams for the online DKE already seen in the lower left panels of Figures 2a and 2b. The lower rows show our measures of metacognitive insight. Insight was generally poorer for the Memory task than for the Movement task, with higher JNDs, indicating lower sensitivity to own performance, and higher SETs, indicating more lax criteria for self-estimation. Participants thus had less insight into their performance in the Memory task, probably because this task was even more cryptic than the Movement task in terms of available feedback.[3]

Crucially, in both tasks, participants in the top tertile showed sharper sensitivity to their performance (lower JND) and a SET which was both more conservative and better calibrated to the objective criterion for success (response error of six pixels or less). The lower tertile of performance included some participants with good metacognitive insight, but also featured some participants who were very insensitive (high JND) and/or had a very lax criterion (high SET). Poorer performance is therefore associated with poorer metacognitive insight in both tasks, consistent with the hypothesis that the DKE arises from metacognitive differences between more and less skilled participants (Kruger & Dunning, 1999).

Does metacognitive insight mediate the DKE?

A causal role for metacognitive differences cannot be tested directly, because the observed relationships are correlational; but we can test for a pattern of mediation that would support a causal role. The DKE is represented by the zero-order correlation between baseline performance and online estimation error. We re-assessed this relationship as a series of semi-partial correlations, controlling baseline performance for variance shared with SET and/or JND (Table 2). The proportion of the squared zero-order correlation accounted for by the

[3]The high number of participants excluded from the Memory task (n =37), due to a failure to fit a significant logistic regression, could also reflect generally less insight in this task.

inclusion of both psychophysical variables was .92 for the Movement task and .77 for the Memory task. The effective mediating variable in the Movement task was SET, whilst JND and SET both made a contribution in the Memory task.

	Movement task (n=80)	**Memory task (n=62)**
Full correlation of baseline performance and online estimation error	-.30	-.58
Semi-partial correlations:		
controlling SET	-.05	-.34
controlling JND	-.25	-.33
controlling SET & JND	-.09	-.28

Table 2. *Full (zero-order) correlation between baseline performance and online estimation error, and semi-partial correlations controlling for variance shared between baseline performance and SET and/or JND. All analyses were performed on ranked data. The significance threshold would be ≥ .22 for the Movement task and ≥ .26 for the Memory task (two-tailed, uncorrected, alpha = .05).*

Experiment 1: Discussion

This first experiment replicated all essential features of the DKE for two novel tasks, of movement and memory. This extends the generality of the DKE, suggesting that it is a near ubiquitous pattern for tasks that are neither trivially easy nor unreasonably difficult (i.e. not performed at floor or ceiling), and for which the available feedback is sufficiently cryptic to leave some uncertainty of self-estimation. In these novel tasks, we replicated prior observations that the DKE is inflated if the index of performance is drawn from the same trials in which estimation error is measured, presumably due mainly to regression to the mean. We also confirmed that the pattern is stronger for relative than for absolute global estimates. However, the effect does not depend on uncertainties in making global estimates, because it is also replicated with a series of online reports of perceived success or failure in individual trials.

Metacognitive insight was generally better for the Movement task, in which there was sensorimotor feedback associated with Movement, than it was for the Memory task, which provided no external feedback. In both tasks, metacognitive sensitivity and calibration were robustly related to actual performance (Table 1), and the relationships remained significant when using the baseline measure of performance from a distinct set of trials (Figure 3). At least some low performers had very poor metacognitive sensitivity (high JND), whereas high performers had generally good sensitivity. Similarly, some low performers had very lax standards for reporting a hit (high SET), whilst high performers were relatively conservatively calibrated. These more conservative criteria were closer to the (non-arbitrary) objective criterion for success, so it seems reasonable to suggest that the self-estimations of high performers were not just more conservative, but better calibrated in absolute terms.

The critical question is whether these metacognitive differences induce the ubiquitous negative correlation between task skill and estimation error (i.e. the DKE). The present experiment cannot test for a causal role of metacognition, but did show a statistical pattern of mediation consistent with a causal relationship. In the Movement task, the mediation was almost complete, with the online DKE reduced from -.30 to -.09 once metacognitive differences were controlled for. In the Memory task, the mediation was partial but substantial, with the online DKE reduced from -.58 to -.28. Experiment 1 thus supports a causal, but not exclusive, role for metacognitive differences in generating the DKE, and suggests that calibration (SET) is particularly influential. It would thus seem that the unskilled are indeed

less sensitive to their own performance, and depart further from reality in their metacognitive calibration, and that this substantially determines their estimation error, as originally hypothesised by Kruger & Dunning (1999).

However, as discussed in the Introduction, the measurement of estimation error may be subject to a performance artefact (over and above regression to the mean), such that low performers are more prone to overestimate just because they have relatively more errors to mistake for successes (e.g. Krajč & Ortmann, 2007). This artefact would entail that estimation errors are negatively correlated with main block performance. The measure of task skill from baseline performance is also related to main block performance (which is simply to say that our tasks have reasonable test-retest reliability). Baseline performance could thus inherit an artefactual negative correlation with estimation error through its positive relationship with main block performance, providing an additional mechanism to drive the DKE. Moreover, our metacognitive variables are also strongly related to main block performance (see Table 1), so some portion of the proposed performance artefact might get misattributed to metacognitive factors in our mediational analysis, leading us to over-estimate their causal role.

These considerations highlight a confound that has received little discussion in the DKE literature. The original hypothesis for the effect is that it follows as a function of task skill, because task skill determines metacognitive insight for a task. However, the measurement of metacognitive insight, whether through estimation errors or through a psychophysical analysis as in the present study, has only ever been done in situations in which high and low skill participants differ in their success at the task. This confounds task skill (ability for a class of task) with task performance (level of success at current instance of task), so does not allow us to disentangle skill- and performance-driven effects. This might seem a subtle distinction, but it could be crucial to a correct understanding of the DKE. If a performance artefact exists such that a high rate of errors boosts overestimation, and vice-versa for a low rate of errors, then to study the effects of task-skill uncontaminated by this artefact, we should really compare estimation errors between high and low skill participants when they perform the task at equivalent rates of success.

A similar argument applies to our psychophysical measures of metacognition. We have implicitly assumed that these reflect relatively stable characteristics, which differ between people with different levels of task skill. But imagine instead that metacognition is

more-or-less modulated by the current level of task performance. In particular, rather than having a fixed criterion for success (SET), a person might adopt a more conservative criterion when objectively more successful, and a more lenient criterion when less successful. In plain terms, we might expect much from ourselves when a task is easy, but give ourselves more leeway when conditions are difficult. For instance, in a general knowledge quiz, we might be satisfied only with exact answers in our specialist area, but happy with strong hunches for questions outside of our expertise. When aiming for a dot, we might want to land comfortably inside the boundary of a big dot, but be happy to clip the outside edge of a smaller dot. Again, one way to disentangle the effects of task skill from those of performance *per se* would be to study the predictive effects of task skill after differences in performance have been eliminated.

This is the purpose of Experiment 2. We set out to test whether specific correlations between baseline performance and three measures of online self-estimation for the Movement task of Experiment 1 would be replicated if between-participant differences in performance (hit rate) in the main block were experimentally eliminated. That is, we tested whether these relationships were driven by task skill, or by task performance.

Experiment 2: Methods

The critical correlations from Experiment 1, under replication in Experiment 2, were between baseline performance and online estimation error (ρ = -.30), baseline performance and SET (ρ = -.26), and baseline performance and JND (ρ = -.37). Our methods and predictions were preregistered on the Open Science Framework https://osf.io/ccgsz/. We report how we determined our sample size, all data exclusions, all manipulations, and all measures. Only the Movement task was used in Experiment 2. All methods were exactly as for Experiment 1, except in the details described below.

Participants and power

Our plan was to calculate a Bayes factor after every ten participants (or nearest break point in testing) using the replication test for correlation developed by Wagenmakers, Verhagen & Ly (2016), to test between the hypothesis that the previously observed correlation was replicated and the null hypothesis of no correlation.[4] Our primary stopping rule was to terminate data collection at the point that the Bayes factors for all three target correlations were sensitive, according to the cut-offs suggested by Jeffreys (1939) (i.e. greater than 3 or less than 1/3). Our secondary stopping rule, defined on frequentist grounds, was that we would stop data collection after 88 valid datasets, if the primary stopping condition had not been met. With a one-tailed alpha of .05 (because the direction of correlation is known), this would provide high power to replicate the correlation with online estimation error (power .90 for ρ = -.30 at n=88), and with JND (power .98 for ρ = -.37 at n=88), and adequate power to replicate the correlation with SET (power .80 for ρ = -.26 at n=88) (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007).

In fact, our primary stopping rule was met after 81 participants had been tested. These participants were students of the University of Edinburgh, with a median age of 20 years (min 18, 1st quartile 19, median 20, 3rd quartile 23, max 32 years), and mostly female (60 female,

[4] This replication Bayes factor was developed for Pearson correlations. We apply it in the present case to Spearman correlations, which are identical with Pearson correlations for ranked data.

21 male) and right-handed (74 right handed, 7 left-handed, by self-report). After exclusions, the final sample had 75 participants.[5]

2.1. Procedure

The procedure was the same as for the Movement task of Experiment 1, except as stated here. In order to try to minimise the number of participants excluded because of time-errors, any participant with an excess of time-errors in the baseline block was given a second attempt at this block. Participants were excluded for time-errors only if they produced an excess (>50) at two attempts of the baseline block, or an excess (>100) in the main block. In each case, the block was discontinued as soon as the maximum number of time-errors was exceeded. In practice, no participants were excluded because of time-errors.

The main block was identical to the baseline block except that, rather than presenting a fixed set of dot sizes, the dot size varied from trial-to-trial. The initial radius, used on the first trial, was set to the median of the absolute deviation from the centre of the dot in the baseline block, rounded to the nearest pixel. Thereafter, each time that the participant hit a dot, the radius decreased by two pixels at the next trial; and each time the participant missed a dot, the radius increased by two pixels at the next trial. This simple adaptive rule was used to titrate the hit rate for each participant in the main block to around 50%. We expected high skill participants to be presented with generally smaller dots (a more difficult task), and low skill participants to be presented with generally larger dots (an easier task), and for all participants to have a similar level of success (~ 50%).

As in Experiment 1, after every valid trial in the main block, the participant had to report whether they thought their response was a hit or a miss. Because we were interested specifically in online self-estimation, we did not include any prospective self-estimates before the main block. However, we did collect retrospective global self-estimates, as a no-cost add-on. A first rating screen asked for an absolute estimate with the wording: “MOVEMENT TASK: IN THE LAST BLOCK, YOU HAD HALF A SECOND TO REACH AND HIT THE DOT. ON WHAT PERCENTAGE OF TRIALS DO YOU THINK YOU HIT THE DOT?”. A second rating screen then asked for a relative estimate: “OUT OF 100 HEALTHY

[5]One participant was excluded because performance in the main block did not fall within the required bounds, and five participants were excluded due to a failure to fit a significant binomial logistic regression.

ADULTS DOING THS TASK, HOW DO YOU THINK YOU WILL COMPARE, IF 0 IS WORST AND 100 IS BEST?"

Data treatment and dependent variables

The data treatment was identical to that in Experiment 1, except that the binomial logistic regression of online self-estimation reports included dot radius as a predictor in addition to response error. This was done because there was a potentially wide variation of dot sizes, and we wanted to account for any possible biasing influence of dot size itself (e.g. the participant might be more likely to report a hit simply because the target was bigger).[6] SET and JND were calculated from the two-factor regression equation, for a dot radius of 14 pixels (the average dot size in Experiment 1).

Participants were excluded at the analysis stage if the titration of performance levels failed, which we operationally defined as a hit rate in the main block below 45% or above 55%. One participant was excluded on these grounds (hit rate 55.5%). We also excluded participants if the binomial logistic regression showed a multicollinearity problem, indicated by a variance inflation factor exceeding four, or if they had fewer than ten estimation responses available in either category (hit or miss), or if the regression did not find a significant effect of response error on self-estimated performance, as assessed by the Wald test for that predictor ($p > 0.05$). In these cases, the psychophysical measures SET and JND could not be meaningfully estimated; five participants were excluded on these grounds.

[6]Dot size could also have been included as a predictor in Experiment 1, but we did not plan to do this, because pilot data had indicated that there was no advantage to doing so. The analyses reported for Experiments 1 and 2 thus adhere to the preregistered plans, but the outcomes would not be meaningfully changed by including dot size as a predictor in Experiment 1, or by not including it as a predictor in Experiment 2.

Experiment 2: Results

The top left panel of Figure 4 shows the successful titration to ~50% hit rate in the main block. The bottom two panels of Figure 4 show that, despite this levelling of performance, SET and JND have similar relationships to baseline performance as in Experiment 1 (cf. Figure 3). The correlation for SET was -.22 (vs. -.26 in Experiment 1), and the correlation for JND was -.38 (vs. -.37 in Experiment 1). For SET, the replication BF_{10} was 4.90, corresponding to 'substantial' evidence for replication, and for JND the replication BF_{10} was 332.30, corresponding to 'extreme' evidence for replication (Jeffreys, 1939). These outcomes indicate that the psychophysical differences in metacognitive insight are driven by task skill, rather than by performance in the current instance of a task.

However, the top right panel of Figure 4 shows that, despite the replication of these metacognitive differences, the DKE itself was not replicated. 'Substantial' evidence was instead found for the null hypothesis of no correlation between baseline performance and online estimation error (replication $BF_{10} = 0.19$). Figure 5 displays the mean estimation error for each tertile of baseline performance, for online estimation and also for the retrospective global ratings, confirming that the DKE was uniformly abolished by the levelling of main block performance. At face value, these data seem to undermine the idea that metacognitive differences could cause the DKE, because these differences are present, but the DKE is not. More precisely, however, the finding shows that the metacognitive differences are *not sufficient* for the DKE, which does not mean that they cannot contribute to its generation. However, it does imply that the causal antecedents of the DKE are more complex than being due simply to metacognitive differences or to performance artefacts, or even their additive combination. We now explore this in more detail in Experiment 3.

Experiment 3

Experiments 1 and 2 suggest that metacognitive differences contribute to the DKE, yet are neither necessary not sufficient for it. Rather, the DKE pattern may be shaped by interactions between differences in metacognitive insight related to task skill, and performance artefacts affecting the measurement of estimation error. Experiment 3 reports some basic statistical modelling to help formalise and visualise these interactions. The task that our simulated participants are performing is arbitrary, provided that they perform neither at floor nor ceiling, that they are neither completely certain nor completely uncertain of their success, and that each response can be conceptualised as having some quantifiable degree of error.

Simple simulation

Assuming that performance errors for a given task are normally distributed, we simulated idealised error distributions for individual participants. On the unit-less scale of response error, the objective criterion for success was set at zero (i.e. errors less than or equal to zero were hits, and errors greater than zero were misses). Each distribution was given the same standard deviation (1), and errors were expressed in terms of standard deviations from the objective criterion. We simulated error distributions with a low mean (-1), a medium mean (0) or a high mean (1). This might represent high, medium and low performers on the same task, or it might represent the same participant performing easy, medium and hard tasks.

Given that participants have some insight into their performance, we simulated psychophysical functions for the probability of reporting a hit across an error range of -5 to 5. All functions were initially given the same JND (0.5). We used three different values of SET, such that the 50% probability of reporting a hit was at a low error (-1), a medium error (0) or a high error (1), akin to participants with an overly-conservative (negative) criterion, a perfectly calibrated criterion, or an overly-liberal (positive) criterion. The three levels of SET were crossed with the three levels of performance, to produce nine simulated datasets (Figure 6). The density distribution of errors is shaded according to the proportion of hit reports (dark grey) and miss reports (light grey) at each level of response error. The black vertical line marks the objective threshold; errors to the left of this line are hits and errors to the right are misses. Underestimation is thus represented by the light grey area to the left of this line, and overestimation is represented by the dark grey area to its right. The estimation error is given

by the difference (overestimation – underestimation), expressed as a percentage of the total area under the curve.

Several observations can be made. First, for a fixed level of task performance (each row in Figure 6), participants with a negative criterion (SET < 0) tend to underestimate, and those with a positive criterion (SET > 0) tend to over-estimate, and the magnitude of estimation errors depends on how well-calibrated SET is to the objective criterion. Second, for a given psychophysical function (each column in Figure 6), high performers tend to underestimate, and poor performers tend to over-estimate, and these differences are more pronounced when self-estimation is less well calibrated to the objective threshold. The performance-driven part of this pattern greatly affects the measurement of estimation error, but needs no explanation in psychological terms, since it is an artefact of a sampling bias: poor performers make more errors, so have they more chances to overestimate, and vice-versa for high-performers.

Extended simulation

We extended our simulation to explore the influence of SET across a wider range, and to incorporate the influence of JND. We simulated datasets combining three levels of performance/task difficulty (mean error 1, 0, -1), and three levels of JND (0.5, 1, 1.5; representing decreasing sensitivity to trial-to-trial variations in error) across a wide range of SET (-4 to 4). For each dataset, we calculated the estimation error for that combination of task performance and psychophysical function. The main patterns are shown in Figure 7.

Figure 7 supports some further observations. At a fixed level of performance (i.e. in each panel), estimation error increases as a sigmoid function of SET, with the sign of SET very largely determining the sign of estimation error. The influence of SET is moderated by JND, with a low JND (i.e. high sensitivity) amplifying the influence of SET, and a high JND dampening it. The only point at which JND has no influence is when SET is coincident with the mean of the error distribution. Although not so apparent in Figure 7, JND similarly moderates the influence of performance error, and its influence differs according to the sign and value of SET (i.e. a complex three-way interaction). These moderating effects of JND are subtle compared with the dramatic effects of SET and performance.

Experiments 2 & 3: Discussion

Experiment 2 found that the metacognitive differences between high and low performers are related directly to task skill, independent of current level of performance, because these differences persisted unchanged when success in the main block was levelled (at ~50%). This supports the broad idea that metacognitive insight may depend, to some extent, on the same core competencies that constitute task-skill (Kruger & Dunning, 1999). A similar result was not found for the DKE itself, which was statistically eliminated by the removal of performance differences in the main block. This implies that the patterns of over- and underestimation that define the DKE are not a direct function of task skill, but are shaped by the current performance of the task.

This was further explored via the simulations of Experiment 3, which combine a Gaussian error distribution (performance) with a psychophysical function (metacognition). This is a simple but general model, applying in principle to any task in which the degree of error can be conceived as varying, and in which participants have some (imperfect) insight into their performance. By varying performance (i.e. mean of the error distribution), alongside metacognitive sensitivity and calibration (i.e. JND and SET), we confirmed that metacognitive factors are neither necessary nor sufficient for the DKE. They are not necessary because the DKE pattern can arise from performance differences, provided only that SET and JND are non-zero; and they are not sufficient, because even if metacognitive differences exist between participants, their effects on estimation errors are modulated by performance (see Figure 7).

The simulations also suggest that, for a fixed level of performance, metacognitive differences should directly determine estimation error (e.g. within any row in Figure 6). However, in the empirical Experiment 2, metacognitive differences did not induce the DKE, despite the fact that performance was fixed at ~50% across participants. To resolve this apparent discrepancy, it is important to note that our performance-levelling method controlled hit rate, but could not ensure equivalence of the underlying error distributions. These may still have varied unpredictably between participants, limiting or masking the influence of metacognitive differences. Unlike our simulations, real participants differ not just in their mean error, but in the shape of their error distribution, for instance in variability, kurtosis and skew, which will cause further perturbations of estimation error.

Our simulations also imply that task difficulty should affect estimation errors. The effect of task difficulty is interchangeable with performance: an easy task is represented by a leftward shift in the error distribution, and a hard task by a rightward shift. The performance artefact predicts more overestimation for the hard task relative to the easy task, as a consequence of these shifts. This is consistent with the prevailing patterns of absolute estimation errors for the small dot (hard task) and large dot (easy task) in Experiment 1 (Figure 2), and also with a large body of data in the wider literature on self-estimation (for a synthesis and theoretical account, see Moore & Healy, 2008). However, whilst people tend to overestimate their *absolute* performance on hard tasks, and to underestimate it on easy tasks; these patterns typically reverse when estimating themselves *relative* to others, giving rise to the better-than-average effect for easy tasks, and worse-than-average effect for hard tasks (Moore & Healy, 2008; Burson et al., 2006; Kruger, 1999). This reversal for *relative* estimates was not so clearly observed in Experiment 1 (Figure 2); rather, the relative estimates were quite similar for small and large dots. This is probably because the relative estimates for the two dot sizes were made in immediate succession. To have systematically estimated different percentiles for the different dot sizes would have required people to reason quite explicitly that task difficulty would affect them differently from everybody else.

General discussion

Since the seminal paper establishing the DKE, there has been debate over whether the effect derives from metacognitive differences between skilled and unskilled people (Kruger & Dunning, 1999), from general biases of self-estimation (Burson et al., 2006; Krueger & Mueller, 2002), or from statistical artefacts (Feld et al., 2017; Krajc & Ortmann, 2008; Krueger & Mueller, 2002). Despite vigorous defences of the metacognitive account (Dunning, 2011; Ehrlinger et al., 2008; Kruger & Dunning, 2002; Schlösser, Dunning, Johnson, & Kruger, 2013), and its enthusiastic dissemination through the wider culture, unambiguous evidence has not been provided, and the debate has persisted. The present paper offers a resolution, showing that each of these factors can contribute to the typical DKE, and further describing how they may interact to determine estimation errors. Our studies employ novel tasks, and online self-estimation. In drawing our main conclusions, we assume meaningful generalisation from these methods to other tasks; and we consider the likely limits of this assumption.

In both Movement and Memory tasks, poor performers did indeed show worse insight, having lower sensitivity to variations in their performance, and more lax standards for success, being willing to claim a hit even when the spatial error was large. This pattern was found in some, though not all, poor performers, whilst high performers uniformly showed high metacognitive abilities, being more sensitive, and having stricter standards, better calibrated to the objective criterion. These metacognitive differences correspond quite well with Kruger and Dunning's (1999) original concept of a ‘metacognitive deficit’ amongst the unskilled. Metacognitive differences were found to partially mediate the DKE, consistent with a causal role in its generation. However, the effect size that these differences could account for is much smaller than the DKE that is widely depicted (e.g. by the famous Figure 1 of Kruger & Dunning, 1999); and other biases can strongly influence and inflate the effect.

The DKE is inflated if the measure of performance is drawn from the same trials as used in the calculation of estimation error. In Experiment 1, the correlations between performance and estimation error were smaller in magnitude by around ~ .3 for the Movement task and ~ .2 for the Memory task, when using a measure of performance from separate (baseline) trials (see Figure 2). Most of this discrepancy is almost certainly due to regression to the mean (Ackerman et al., 2002; Burson et al., 2006; Feld et al., 2017; Krueger & Mueller, 2002; Nuhfer et al., 2016, 2017). The DKE is also stronger for relative than for

absolute self-estimates, probably due to regressive effects associated with the added uncertainty about how other people perform (Moore & Healy, 2008). The combined effect of these two methodological factors can be dramatic. In the Movement task of Experiment 1, the DKE was represented by a correlation of -.92 when relative estimates and main block performance were considered, yet receded to non-significant levels (as weak as -.07) for absolute estimates and baseline performance.

Our main operational measure of the DKE in these studies was the correlation between task-skill (baseline performance) and online estimation error. In both tasks, DKE magnitude was reduced by about .3 by controlling for metacognitive differences. However, this apparent mediation may overstate the importance of metacognition, because metacognitive measures are confounded with level of performance. Even if low and high performers had exactly the same metacognitive skills, low performers might overestimate more often just because they have more failures about which to be mistaken (Krajc & Ortmann, 2008). In Experiment 2, when we levelled the performance across participants, the metacognitive differences between more and less skilled people persisted, but the DKE itself was abolished. Thus, although the metacognitive differences suggested by Kruger & Dunning (1999) are real – at least for some tasks – they are not sufficient for the DKE. Considering that other biases can also induce the pattern, we must conclude that metacognitive differences are not necessary either.

Metacognitive factors do matter, but so does the distribution of performance errors. We explored the interaction of these influences via the simulations in Experiment 3. Metacognitive sensitivity, other than being a requirement for a valid psychophysical function, proved to be a minor influence, modulating the powerful effects of metacognitive calibration and performance. Metacognitive calibration is the principal determinant of the sign of misestimation: when the subjective criterion is higher than the objective criterion, there is overestimation, and when lower, there is underestimation. Overestimation further increases as the subjective criterion becomes more liberal, or as objective performance decreases, and these effects of calibration and performance amplify one another. It may not be very meaningful to try to isolate an exact portion of the DKE that is due to metacognitive factors, because estimation errors depend in complex ways upon the particular mix of metacognitive sensitivity and calibration, and the particular distribution of response errors. Estimation error is certainly not a direct read-out of metacognitive skill.

If this is true for our Movement and Memory tasks, then it may be even more so for higher intellectual and social tasks, where the task itself may be more difficult to define, the size of the response error more ambiguous, the objective criteria for success more opaque, and the feedback more cryptic. (For example, one of the tasks originally studied by Kruger and Dunning was judgement of humour.) Furthermore, when making global self-estimates, rather than trial-by-trial judgements, uncertainty can only increase, especially if required to rank oneself relative to unknown others. The factors and biases inducing over- or under-confidence is the subject of a large literature, beyond the scope of this paper (see Moore & Healy, 2008, for a useful synthesis). Nonetheless, we suggest that metacognitive insight is highly unlikely to be a *more* direct determinant of estimation error in these more complex scenarios. In general, whilst metacognitive differences linked to task-skill may contribute to the DKE, they do not ensure it, and the DKE does not imply underlying metacognitive differences. If the aim is to study metacognitive skill, then researchers should strive to use measures that are free from the many confounds that estimation errors entail (Fleming & Lau, 2014).

Of course, none of our findings cast doubt on the DKE as an empirical phenomenon. On the contrary, our data extend the pattern to novel domains. It is widely true that poor performers overestimate themselves more than high performers, and our data confirm that poor performers may indeed have less metacognitive insight than high performers. But it would be a gross oversimplification to say that poor insight is *the reason* for overestimation amongst the unskilled. At least as much explanatory work is done by performance artefacts, and the pattern can be induced (and is very often inflated) by a host of other factors and biases, some psychologically interesting, and some 'merely' statistical. Shakespeare's poetic depictions of the fool and the wise man thus remain as apt as ever. By contrast, the modern meme that *stupid people are too stupid to know they are stupid* is a misrepresentation, propounded (perhaps) by those who know sufficiently little of the evidence.

Context

Our interest in the Dunning-Kruger effect (DKE) stems from our work on anosognosia for hemiplegia following brain damage. This is a condition in which patients with a weak or paralysed limb seem to be unaware of their disability. When asked to move a paralysed arm, a patient with anosognosia may report that they have done so, despite the fact that they manifestly have not. The DKE has been proposed as "*a psychological analogue to anosognosia*", arising in healthy people who are unskilled in a specific domain (Kruger & Dunning, 2002, p. 1130). We conceived our tasks originally to study self-estimation in clinical populations, but we began by collecting control data, in groups of healthy young and older adults. We quickly became interested in the patterns emerging, and saw the potential value of our methods for testing the role of metacognition in the DKE. These unpublished studies provided rich pilot data for the present experiments, allowing us to optimise our tasks, and to preregister our design, with well-informed power analyses and clear predictions.

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.

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.

Dunning, D. (2011). *The Dunning-Kruger effect. On being ignorant of one's own ignorance*. *Advances in Experimental Social Psychology* (1st ed., Vol. 44). Elsevier Inc.

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.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. A. A.-G. A. A.-G. (2009). Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–60.

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Feld, J., Sauermann, J., & de Grip, A. (2017). Estimating the relationship between skill and overconfidence. *Journal of Behavioral and Experimental Economics*, *68*, 18–24.

Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, *8*(July), 1–9.

Hall, M. P., & Raimi, K. T. (2018). Is belief superiority justified by superior knowledge? *Journal of Experimental Social Psychology*, *76*, 290–306.

Jeffreys, H. (1939). *The theory of probability* (1st ed.). Oxford, UK: Oxford University Press.

Klayman, J., Soll, J. B., González-Vallejo, C., & Barlas, S. (1999). Overconfidence: It depends on how, what, and whom you ask. *Organizational Behavior and Human Decision Processes*, *79*(3), 216–247.

Krajc, M., & Ortmann, A. (2008). Are the unskilled really that unaware? An alternative explanation. *Journal of Economic Psychology*, *29*(5), 724–738.

Krajč, M., & Ortmann, A. (2007). Really That Unaware ? *Centrum*.

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. (1999). Lake Wobegon be gone! The "below-average effect" and the egocentric nature of comparative ability judgments. *Journal of Personality and Social Psychology*, *77*(2), 221–232.

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), 1121–34.

Kruger, J., & Dunning, D. (2002). Unskilled and unaware--but why? A reply to Krueger and Mueller (2002). *Journal of Personality and Social Psychology*, *82*(2), 189–92.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review*, *115*(2), 502–17.

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*, *9*(1).

Nuhfer, E., Fleisher, S., Cogan, C., Wirth, K., & Gaze, E. (2017). How random noise and a graphical convention subverted behavioral scientists' explanations of self-assessment data: Numeracy underlies better alternatives. *Numeracy*, *10*(1).

Schlösser, T., Dunning, D., Johnson, K. L., & Kruger, J. (2013). How unaware are the unskilled? Empirical tests of the "signal extraction" counterexplanation for the Dunning-Kruger effect in self-evaluation of performance. *Journal of Economic Psychology*, *39*, 85–100.

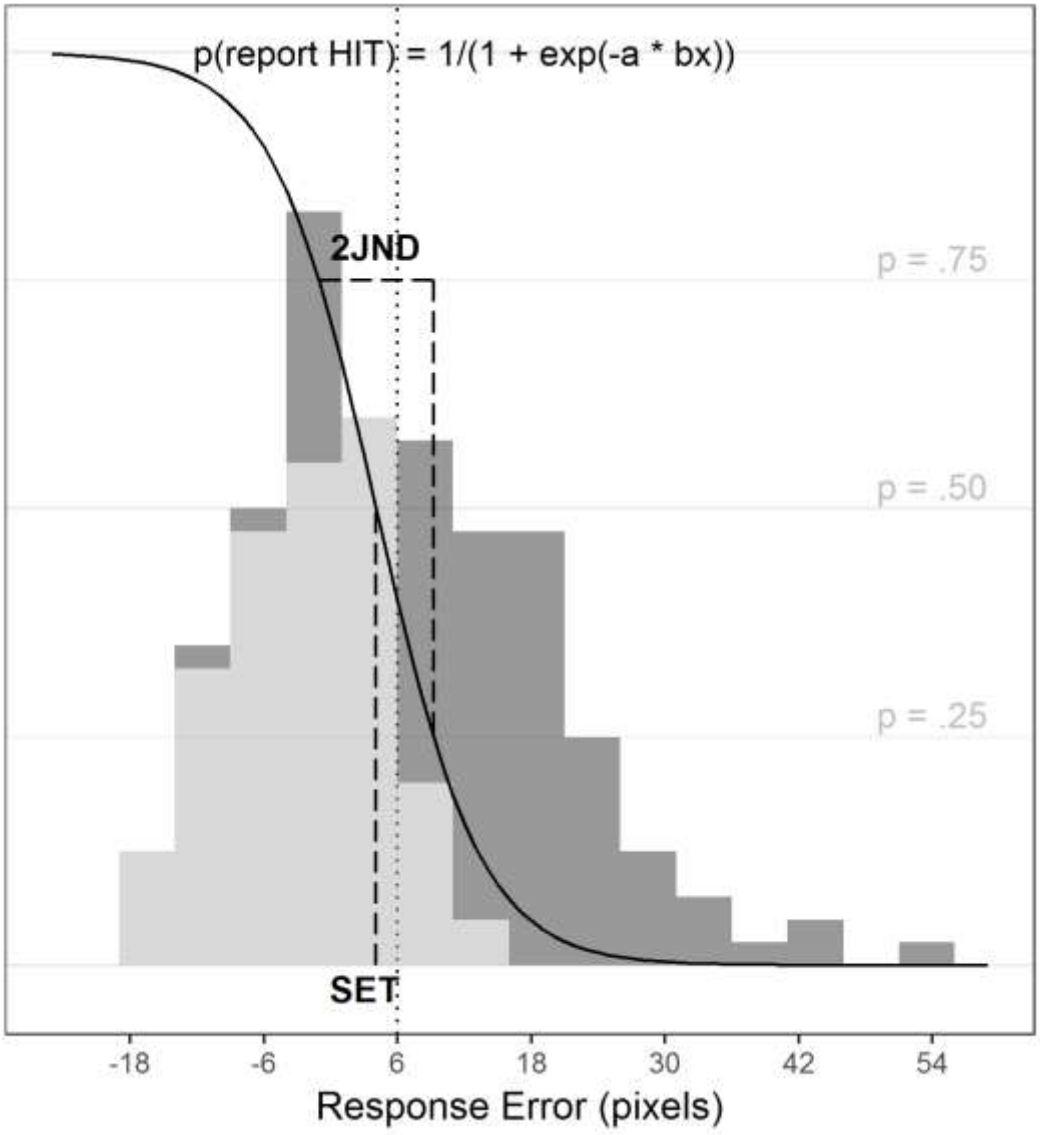

Figure 1. Diagram to illustrate dependent measures for the main block, for one participant performing the movement task. The histogram shows response frequency by size of response error, across 200 pointing movements, with negative errors inside the dot boundary and positive errors beyond the dot boundary. The vertical dotted line indicates the objective threshold for a hit, which includes any responses within a six pixel penumbra around the dot. The percentage of responses falling to the left of this dotted line is the actual hit rate. The histogram is shaded by the frequency with which the participant reported a hit (light grey) or a miss (dark grey). The percentage of (light grey) hit responses is the online estimated hit rate; the online estimation error is then obtained by subtracting the actual hit rate. The black curve is the logistic function relating the probability (p) of reporting a hit to the size of the response error, according to the equation shown. The subjective error threshold (SET), at which the participant is equally likely to report a hit or a miss, is the x intercept at p = .50. The participant's sensitivity to variations in response error is indexed by the just noticeable difference (JND), calculated as (half of) the difference in response error between p = .25 and p = .75. The participant shown here (#52) has an actual hit rate of 58.5%, and an online estimated hit rate of 46.5%, giving an online estimation error of -12.0%. SET is 4.08 pixels and the JND is 5.13 pixels. The regression coefficients are: a = 0.87; b = -0.31.

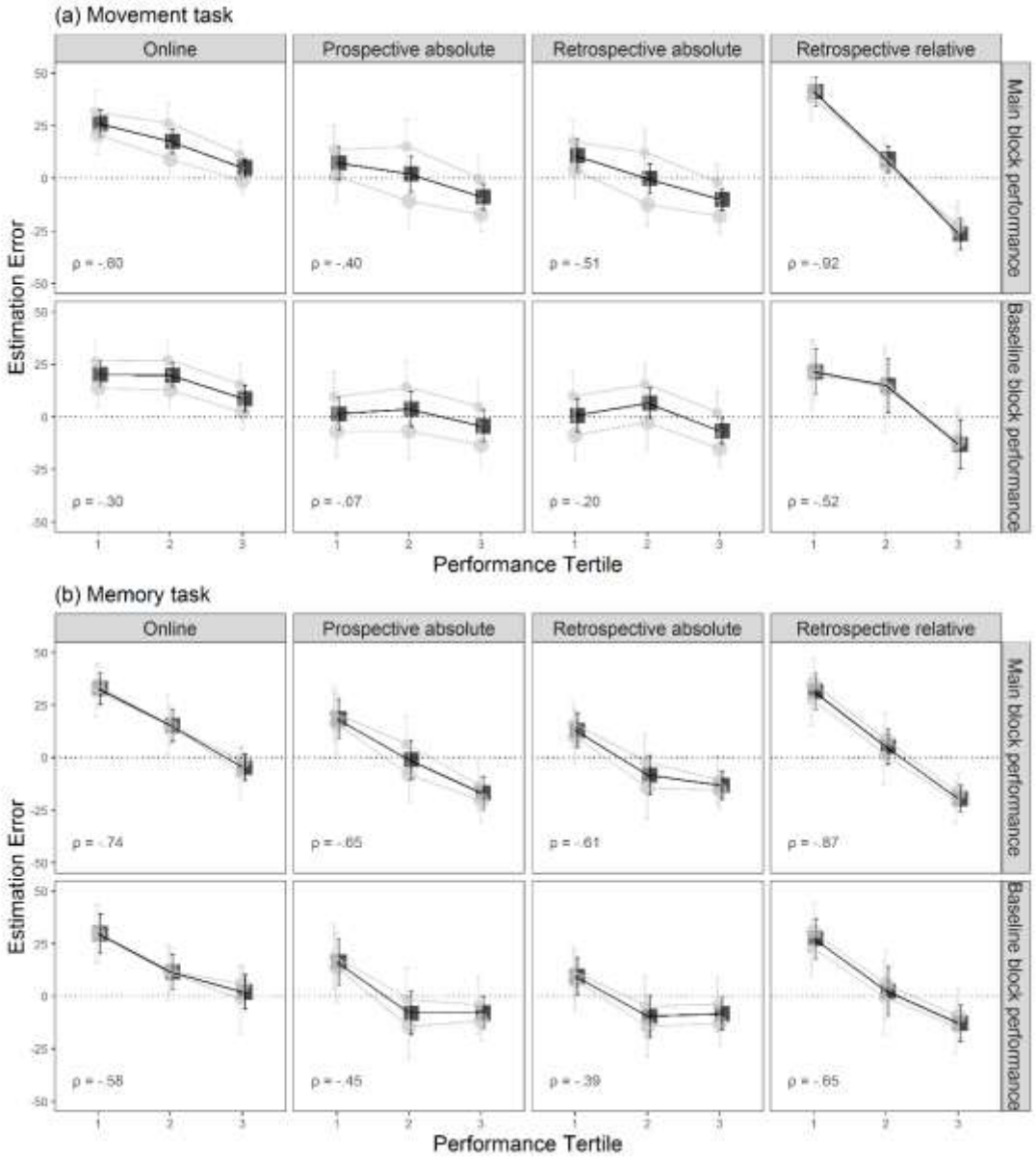

Figure 2. Experiment 1. Relation between performance and estimation error for **(a)** the movement task (n = 80) and **(b)** the memory task (n = 62). Performance is defined by hit rate in the main block (upper row) or the baseline block (lower row). Estimation error is derived from online self-estimation, prospective or retrospective absolute estimates, or a retrospective relative (percentile) estimate. The means are split by performance tertile (where 1 is lower and 3 is upper). Black squares show means (+/- between-subject 95% CIs) across all targets. Small and large grey circles show means (+/- 95% within-subject CIs) for small and large targets respectively. Spearman's rho is reported for each plot, indexing the strength of relation between actual performance and estimation error across all participants.

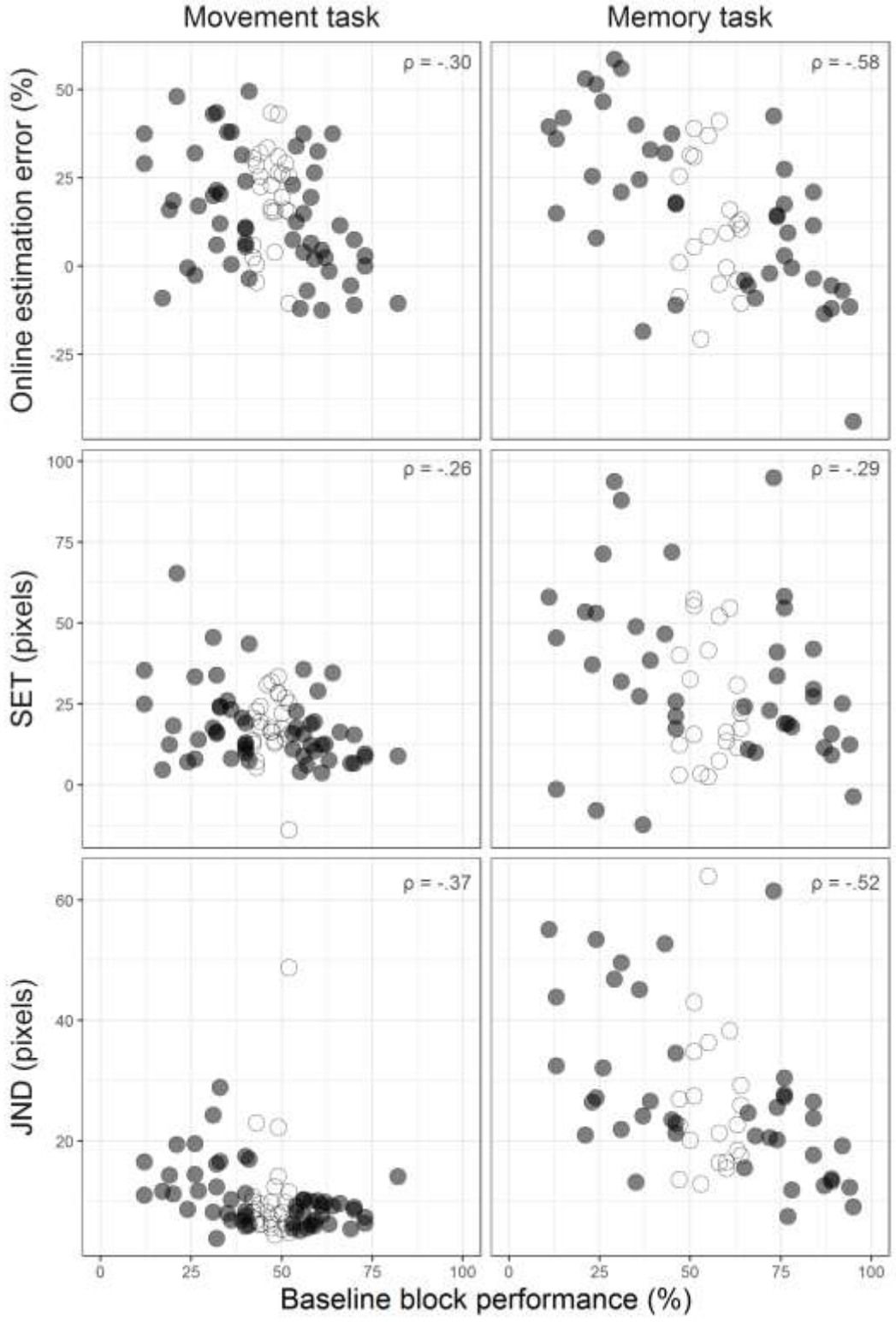

Figure 3. Experiment 1. Relation of baseline block performance (% hit rate) to online self-estimation measures for the main block, for the movement task (n = 80) and memory task (n = 62), with Spearman's ρ for each plot. Participants in the middle tertile of performance are plotted as unfilled dots to visually separate performance tertiles. One outlying participant is omitted from the SET and JND plots for the memory task to avoid compression of the y-axis; this bottom-tertile participant had extremely high values for SET (207) and JND (156).

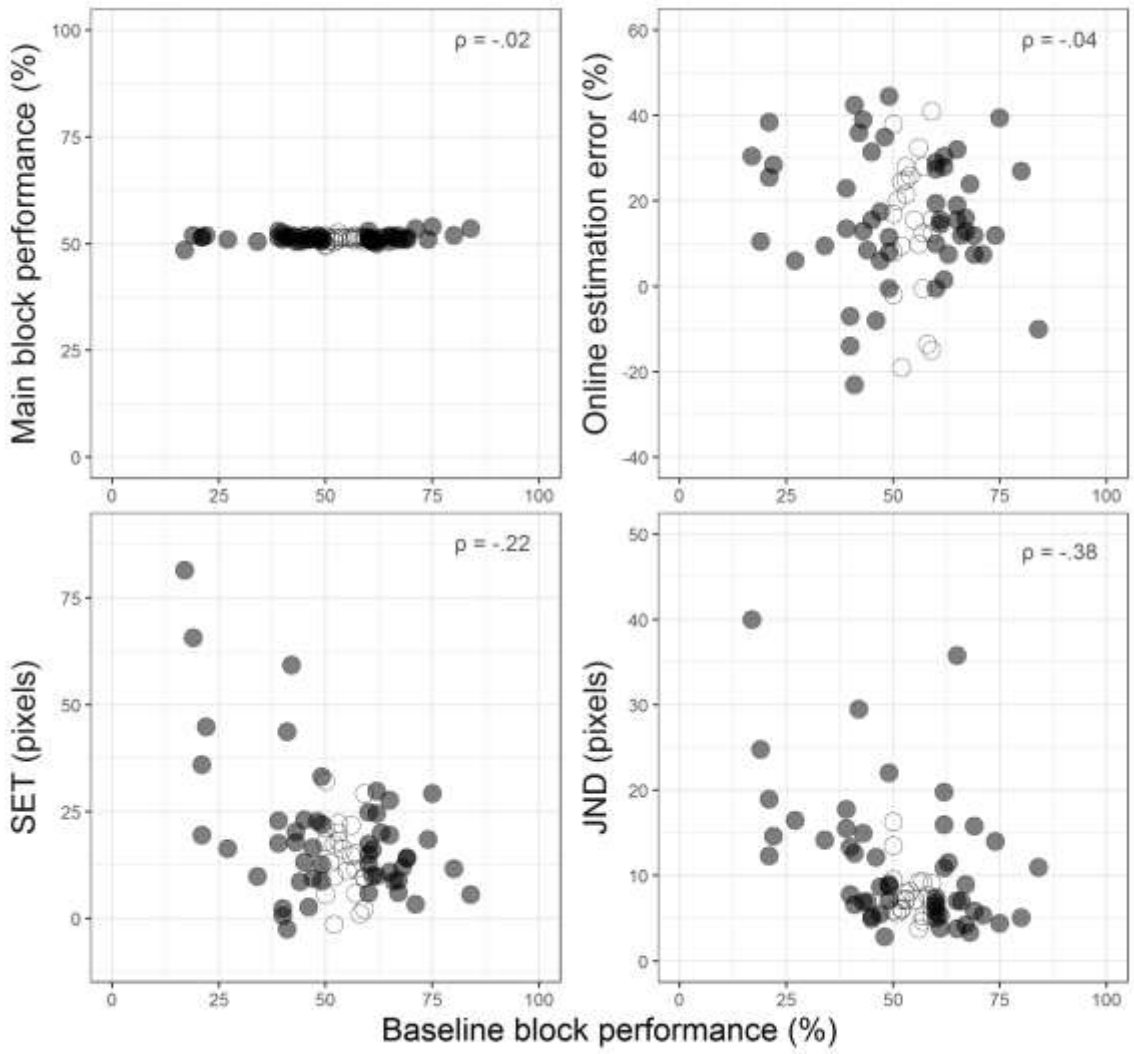

Figure 4. Experiment 2 (n=75). Relation of baseline block performance to (levelled) main block performance, and three online self-estimation measures, with Spearman's ρ for each plot. Participants in the middle tertile of task skill are plotted as unfilled dots to visually separate performance tertiles. The top left plot confirms that the titration of task performance was effective, levelling performance to 50% across the spectrum of task skill. Unlike in Experiment 1, there is no significant relationship between baseline performance and online estimation error, but negative relationships with both SET and JND are replicated.

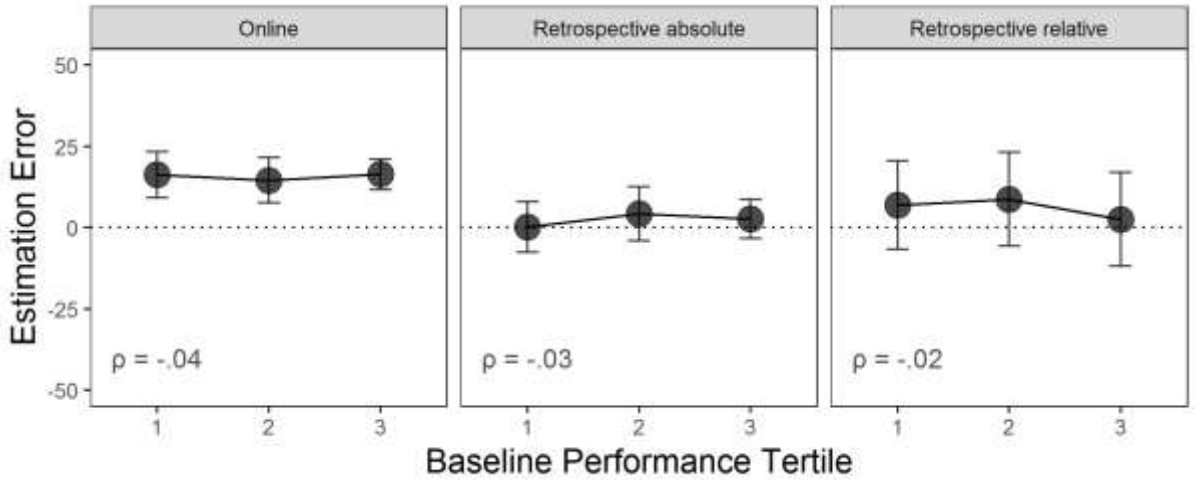

Figure 5. Experiment 2 (n = 75). Relation between baseline performance and estimation error. Estimation error is derived from online self-estimation, a retrospective absolute estimate, or a retrospective relative (percentile) estimate. The means are split by baseline performance tertile (where 1 is lower and 3 is upper). Error bars show between-subject 95% CIs. Spearman's rho is reported for each plot, indexing the strength of relation between baseline performance and estimation error across all participants. The expected DKE pattern, of a negative relationship between baseline performance and estimation error, is absent, due to performance (hit rate) having been levelled in the main block. Note that the relative estimation errors are somewhat arbitrary; they are calculated from the subtraction of actual percentile from estimated percentile but, due to the levelling of performance, the actual percentiles are determined by tiny differences within a compressed range (45-55% hit rate).

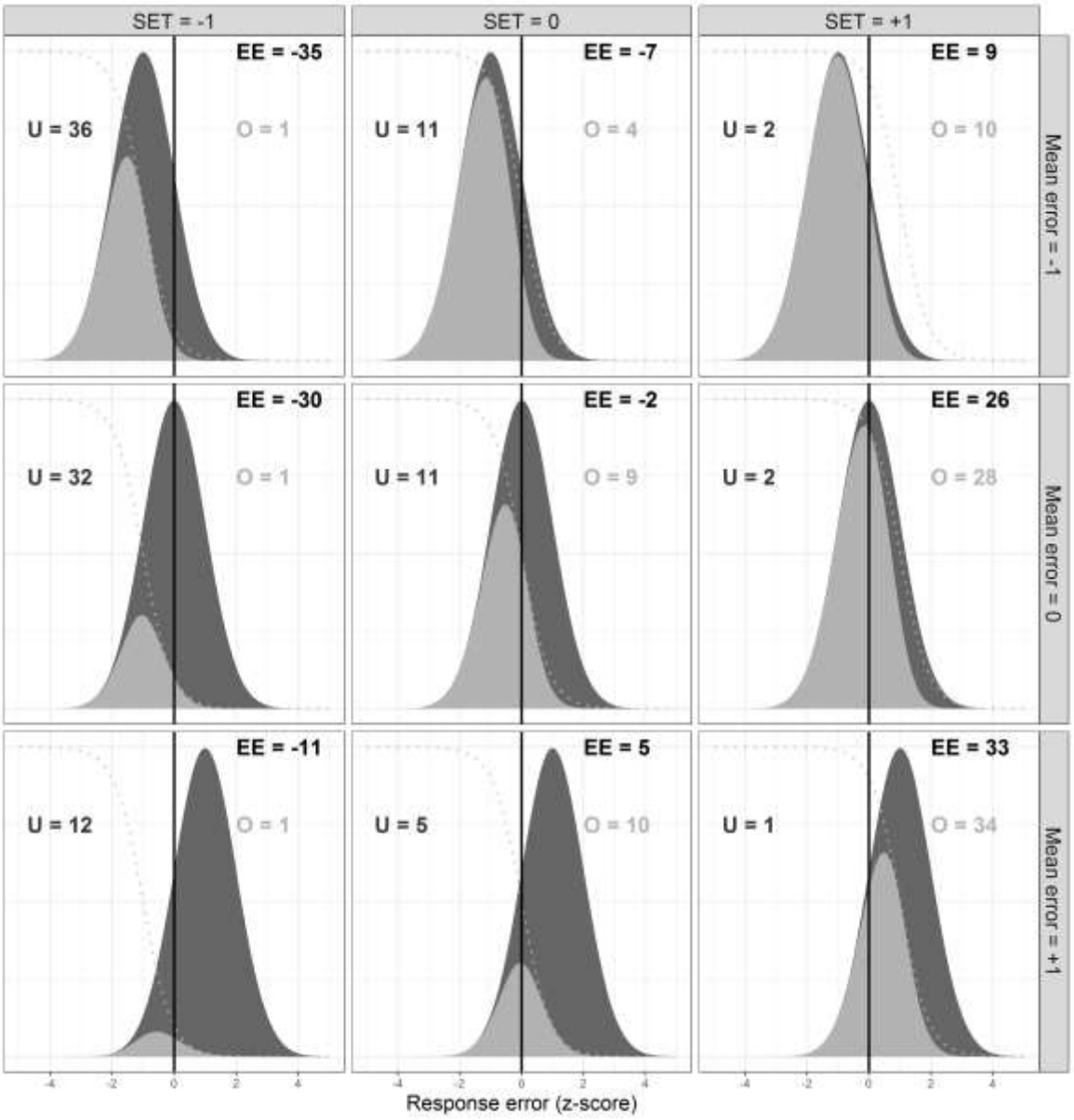

Figure 6. Idealised model of the combined effects of SET and performance error on measures self-estimation. Each plot shows a normal density distribution of errors (SD 1), shaded according to the proportion of hit reports (light grey) and miss reports (dark grey) at each level of error. The vertical black line marks the objective threshold; errors to the left of this line are hits and errors to the right are misses. Underestimation (U) is thus given by the dark grey area to the left of the line; and overestimation (O) by the light grey area to the right of the line. Estimation error (EE) is given by the difference (O-U), with all values expressed as a percentage (to the nearest percent) of the total area under the curve. The nine plots combine three levels of SET (columns) with three levels of mean performance error (rows). The JND is fixed at 0.5 (i.e. the error distance between 75% and 25% hit reports was 1 SD).The grey dotted line shows the psychophysical function relating the probability of a hit report to response error; this function is the same for each plot within the same column. EE varies as both with SET and with mean error.

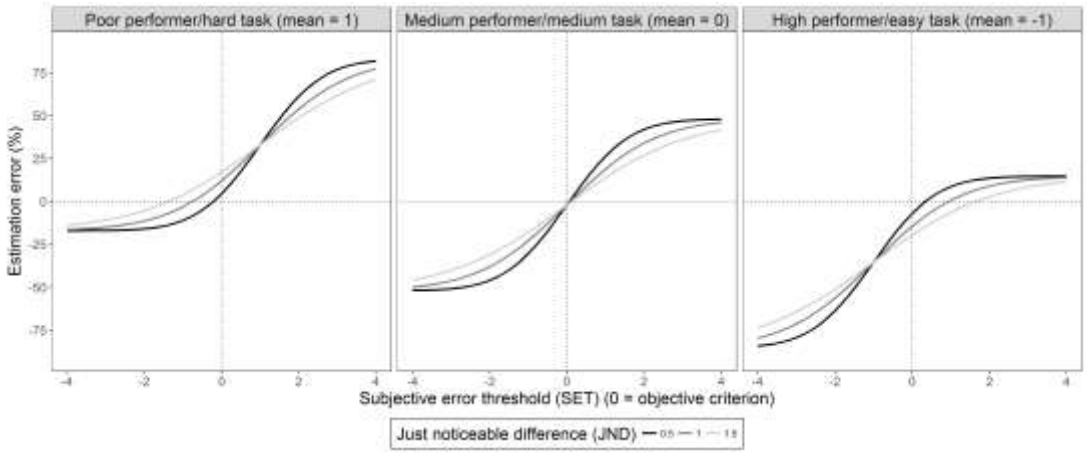

Figure 7. Simulated estimation error as a function of SET and JND. Each panel shows results for a different level of task performance defined by mean error. A higher mean error (left panel) could represent a more skilful participant and/or an easier task, and a lower mean error (right panel) a less skilful participant and/or a harder task.