



THE UNIVERSITY *of* EDINBURGH

Edinburgh Research Explorer

Wise up

Citation for published version:

McIntosh, RD, Fowler, EA, Lyu, T & Della Sala, S 2019, 'Wise up: Clarifying the role of metacognition in the Dunning-Kruger effect', *Journal of Experimental Psychology: General*, vol. 148, no. 11, pp. 1882-1897.
<https://doi.org/10.1037/xge0000579>

Digital Object Identifier (DOI):

[10.1037/xge0000579](https://doi.org/10.1037/xge0000579)

Link:

[Link to publication record in Edinburgh Research Explorer](#)

Document Version:

Peer reviewed version

Published In:

Journal of Experimental Psychology: General

Publisher Rights Statement:

© American Psychological Association, 2019. This paper is not the copy of record and may not exactly replicate the authoritative document published in the APA journal. Please do not copy or cite without author's permission. The final article is available, upon publication, at: 10.1037/xge0000579

General rights

Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Author copy, *Journal of Experimental Psychology: General*, accepted 7 Jan 2019.

DOI: 10.1037/xge0000579

Wise up: clarifying the role of metacognition in 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: An early version of this paper, from prior to peer review, is archived at PsyArXiv, doi: 10.31234/osf.io/czms3. The preregistered plans and materials for this study, and full raw data are archived at the <https://osf.io/ccgsz/>. Full data and analysis code for the final paper are archived at <https://osf.io/8wjck/>

Acknowledgements: We are grateful to Alice Calder and Ink Pansuwan for assistance with data collection, to Adam Moore and Steve Loughnan for comments on an earlier draft of the manuscript, and to Tom Booth for statistical advice on path analysis.

*Corresponding author

Robert D McIntosh

Psychology, University of Edinburgh

7 George Square, Edinburgh, EH8 9JZ

Tel: +44 131 6503444

Fax: +44 131 6503461

© 2019, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final is available, upon publication, via its DOI: 10.1037/xge0000579

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, 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 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 related to task skill, but a path analysis showed that their net contribution to the DKE was weak. The major driver of the DKE was the level of task performance. 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 at equivalent levels of success. Metacognitive measures were again related to task skill but the DKE pattern was eliminated. We present a simple model of these findings, showing that metacognitive differences can contribute to the DKE, but are neither necessary nor sufficient for it. This analysis clarifies and quantifies how metacognitive insight and other factors interact to determine this famous effect.

Keywords: Dunning-Kruger effect; metacognition; performance monitoring; self-evaluation; overconfidence

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 pattern is not in doubt; the basic fact of relatively more overestimation amongst the objectively poorest performers is 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.

Theoretical accounts of the DKE

The original, ‘dual-burden’ 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 more accurate, better-calibrated self-estimates. However, when ranking

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, but show less consistent under- or over-estimation when making absolute estimates (Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008). Low performers 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 pattern. 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). The usual 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 $y-x$ (for a general primer, see Campbell & Kenny, 1999). In concrete terms, as long as participants do not have perfect insight into their performance, 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. It is driven by unreliability in the measurement of performance, so it can be offset statistically by quantifying and controlling for this unreliability (Ehrlinger et al., 2008; Feld et al., 2017; Krueger & Mueller, 2002; Kruger & Dunning, 2002), or it can be avoided entirely 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 attenuated, 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 a 'noise-plus-bias' model, in which uncertainty of self-estimation (noise) is combined with a bias towards a task-specific default estimate. This model assumes that participants lack metacognitive insight into their performance, making their estimates uncertain, but that this is equally true for all participants, regardless of task skill. The model was proposed to account for *relative* self-estimates, so the fact that the DKE arises even for absolute estimates has been used to discount it (see Dunning, 2011). However, it would be possible to extend the noise-plus-bias model to absolute self-estimates, by proposing that people have task-specific biases to estimate a certain level of performance, not just a certain relative standing. If this default level is optimistic (i.e. higher than the true mean), then the poorest performers will overestimate, and the top performers will seem better calibrated, even if all participants are equally lacking in metacognitive insight.

The question of metacognition

Thus, to support the original, dual-burden account of the DKE, it is not enough just to show that the effect persists once regression-to-the-mean has been controlled for, or when absolute self-estimates are used. It is necessary to positively demonstrate metacognitive differences between high and low performers, 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 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.

Foundations for an empirical test

The original hypothesis that the DKE is due to differences in metacognitive insight linked to task ability remains open to debate. Indeed, no study (to our knowledge) has had metacognitive measures that could furnish a fair test of this hypothesis. For this purpose, we require adequate measures of metacognition embedded within tasks that elicit the DKE. The defining character of the DKE is a relative overestimation of own-performance amongst the

least skilled participants; and the range of tasks on which this pattern has been replicated is wide. The majority of these tasks have emphasised high-level intellectual or social expertise, such as logical reasoning, judgement of humour, or domain-specific (or general) knowledge (see Dunning, 2011, for an overview). In principle, though, the same pattern might be expected for any domain of skilled performance, if certain task conditions are met.

First, the task should be within the competence of participants, yet neither too easy nor too difficult, so that between-participant variations in performance can be measured across the range. Second, test-retest reliabilities should be high enough that these variations in performance can be meaningfully linked to differences in task skill. Third, and crucially, feedback should be sparse enough that participants have some uncertainty about their level of success. Dunning (2011, p276) has described this critical condition as the absence of a ‘direct-access cue’: a lack of direct feedback. He pointed out that many intellectual tasks do not have a directly observable outcome, whereas physical tasks often do, which may help explain why correlations of self-evaluation with real ability are often stronger for physical tasks (Mabe & West, 1982). For the intellectual task of judging the comedy value of jokes, for instance, there is no direct feedback; but for the physical task of shooting basketballs, we can just see how each ball lands. However, it is important to note that this is not an intrinsic difference between domains; we could just as well reverse it by occluding the view of the hoop after each ball is launched, and by providing a live comedy audience to give direct feedback on how each *joke* lands. The critical condition is the lack of direct feedback, not the domain of expertise.

For metacognitive differences to drive the DKE, it should also be plausible that insight could draw on some of same processes that underlie task performance. This is almost guaranteed by the condition that direct feedback is unavailable. If we cannot judge our success by observing an outcome directly, our recourse must be to the internal process that informed our response, and/or to indirect cues correlated with it. In a test of knowledge, our certainty about the rightness of our response could be a function of the clarity of our semantic representation, which we may introspect upon directly, or infer from our ease of retrieval, or behavioural correlates such as decision speed (Dunning & Perretta, 2002; Kelley & Lindsay, 1993). Provided – and this is not guaranteed – that our representations are generally clearer for correct responses, then our sense of certainty will roughly track our chance of success, and we will have some metacognitive insight. The overlap between cognitive and metacognitive processes could only ever be partial (by definition, second-order

metacognition cannot be *identical* with first-order cognition), and it may vary from task-to-task. But there is no reason to think it should be specific to intellectual expertise, excluding simpler cognitive and physical skills (Augustyn & Rosenbaum, 2005). In any case, the predicted relationship between skill and insight is an empirical matter. It cannot just be asserted for any task; it needs to be tested, using adequate measures of metacognition.

A novel approach to the DKE

In establishing such measures, we pick up on Burson and colleagues' (2006) conceptual distinction between *sensitivity* and *calibration* as separable sub-components 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 objective criterion. 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 thus the role of metacognition in determining the DKE.

In the present study, we move away from higher-level intellectual skills, to the more tractable domains of movement and spatial memory: pointing at a dot on a screen, or recalling its position after a 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 participant level. The tasks are simple, but they fulfil the core requirements for the DKE to emerge. They are within the competence of participants, yet not at floor or ceiling, and test-retest reliabilities are high enough to establish between-participant differences in skill. We have

implemented the tasks with sparse feedback, so 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. 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, and 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 an aggregate analysis of estimation error, to determine whether the DKE generalises to these novel tasks, and to an online estimation method. 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, dual-burden account of the DKE: that less skilled participants not only perform more poorly, but 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/> (see *Supplemental materials S1*). Experiment 1 was approved by the Psychology Research Committee, University of Edinburgh (approval#105-16171). 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. 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. The basic task was to point to a target dot presented on screen. 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. The purpose of the baseline block was to provide experience

¹Data 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.

of the task, and to obtain an independent measurement of performance. The task experienced in the baseline block was not identical to that in the main block, because different target sizes were used, but it was neither easier nor harder than the main block task, because the average target size was the same.

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. Dot presentation was initiated by the participant depressing the start button. The dot disappeared as soon as the participant released the button to initiate a response, so the participant could not directly observe whether or not their hand landed on target. 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), which onset 450 ms after the initiation of the response (i.e. 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. Prospective global estimates, as well as retrospective estimates, have often been used in the context of the DKE (e.g. Edwards, Kellner, Siström, & Magyari, 2003; Feld et al., 2017; Parker, Alford, & Passmore, 2004; Simons, 2013; Tenenbergh & Murphy, 2005), but they have an additional role in our design. It is possible that our use of online trial-by-trial reporting during the main block could affect retrospective global estimates, perhaps making them more accurate than usual, because of closer attention to performance during the task. The inclusion of

prospective estimates provides at least one global self-assessment that could not be influenced by the online reporting method.

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.

For the psychophysical analysis of metacognitive insight, we fitted a binomial logistic regression to predict the self-estimation report (hit or miss) from the objective response error, for each participant. From this function, we 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 be perfectly calibrated to the objective criterion for a hit). Variations in SET may reflect real differences in where participants perceive the boundary between accurate and inaccurate performance to be, and it will also be affected by more general response biases. For instance, a general (optimistic) bias to report a hit, unless one is certain that the response was unsuccessful, will push the value of SET upwards. Overall, SET can be interpreted as indicating how strict or lenient participants are in assessing their own performance. We also 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.

²Pilot 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.

Figure 1 shows worked examples of this analysis for the Movement task, for one participant with good metacognitive insight, and one with poor insight. Participant 32 (Figure 1a) has a steep transition from hit to miss responses ($JND = 5.55$ pixels), calibrated almost-perfectly to the objective threshold of 6 pixels ($SET = 6.09$ pixels). The transition for Participant 49 (Figure 1b) is much more shallow ($JND = 16.96$ pixels), and miss reports do not reach a majority until the objective error is quite large ($SET = 43.63$ pixels). In these examples, metacognitive sensitivity (JND) and calibration (SET) are both relatively low (Participant 32) or relatively high (Participant 49); but in principle these measures could vary independently. Participants were excluded at the binary logistic regression stage if either type of self-report (hit or miss) was too infrequent to model reliably (fewer than ten reports overall for that category), or if the regression did not find a significant effect of response error, as assessed by the Wald test ($p > 0.05$). In these cases, neither SET nor JND could be meaningfully estimated.³

Our main inferential analyses were based on ranked data (i.e. Spearman, rather than Pearson correlations). 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.

³Three participants were excluded from the Movement task, and 37 participants from the Memory task, by these criteria. See *Supplemental materials S2* for consideration of the influence of these exclusions on our findings.

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 ($\rho = .58$ 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 of estimation, 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 purposes, negative correlations are obtained for online self-estimation (lower left panels of Figures 2a and 2b; $\rho = -.30$ and $-.58$ for Movement and Memory tasks respectively).

Online self-estimation

	Main block	Estimated			
	hit rate	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 $\geq .22$ for the Movement task and $\geq .26$ for the Memory task (two-tailed, uncorrected, $\alpha = .05$). Values exceeding this threshold are shown in bold.

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. 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.⁴

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

⁴See *Supplemental materials S2* for consideration of the influence of these exclusions on our findings.

causal role. The DKE is represented by the negative relationship between baseline performance and online estimation error (lower left panels of Figures 2a and 2b). For each task, we assessed the influence of the variables mediating this relationship, via a path analysis with robust maximum likelihood estimation (Rosseel, 2012).⁵

According to the dual-burden account of the DKE (Kruger & Dunning, 1999), low skill entails poor performance and poor metacognitive insight, which together induce inflated estimation errors. In the present data, hit rate in the baseline block is the measure of task skill, and hit rate in the main block is the measure of task performance, while metacognitive insight is broken down into metacognitive calibration (SET) and sensitivity (JND). We can thus specify three indirect paths mediating between task skill and estimation error: a path via SET and a path via JND, which together give the total metacognitive path; and a third path via task performance. These indirect paths are shown in Figure 4, along with the estimated strength and significance of each.

In both tasks, SET and JND are higher amongst less skilled participants, and mediate the relationship between skill and estimation error (i.e. the DKE), but in opposing directions. Positive metacognitive calibration (high SET) tends to induce overestimation, promoting the DKE, but poor metacognitive sensitivity (high JND) tends to counteract it. The overall mediating effect of metacognition is negative, promoting the DKE, but this net influence is not significant in either task (Movement task $-.13$, $p = .17$; Memory task $-.09$, $p = .31$). The mediation by task performance is stronger, considerably so in the Memory task, and highly significant (Movement task $-.18$, $p < .0005$; Memory task $-.51$, $p < .0005$).

⁵Our pre-registered analysis for this section was based on a series of semi-partial correlations. Following the suggestion of an anonymous reviewer, we have replaced this with a path analysis, which offers richer insights into the indirect paths mediating the DKE.

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, 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 (Burson et al., 2006; Feld et al., 2017; Klayman et al., 1999). We also confirmed that the pattern is stronger for relative than for absolute global estimates (see Dunning, 2011). However, we further showed that 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 than for the Memory task, but metacognitive sensitivity and calibration were robustly related to actual performance in both tasks (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. It would thus seem that the unskilled are indeed less sensitive to their own performance, and unduly optimistic in their metacognitive calibration, as originally hypothesised by Kruger & Dunning (1999).

The critical question is whether these metacognitive differences mediate the negative correlation between task skill and estimation error (i.e. the DKE). A path analysis showed that lax metacognitive calibration (high SET) did promote the DKE, but that this was largely counteracted by an opposing effect of poor metacognitive sensitivity (high JND). Intuitively, a high JND indicates a shallower psychophysical function, which attenuates the importance of the precise point of transition represented by SET (see *Supplemental materials S3*, for a full explanatory model). In both tasks, the total effect of metacognitive factors was weakly to promote the DKE, but was not statistically significant. Task performance (hit rate) in the

main block mediated the majority of the DKE, and this indirect path between task skill and estimation error was highly significant in both tasks.

Given this suggestive pattern, we can ask a deeper experimental question about underlying causes. Specifically, we can seek to disentangle the two parts of the proposed dual-burden of the unskilled. One part is a lack of metacognitive insight, which we have confirmed, but which may play a minor role. The other is simply that performance itself is poor. As noted in Kruger and Dunning's (1999) original report, incompetent individuals are at the bottom of the distribution, so it is nearly impossible for them to underestimate their relative rank (see also Krajč and Ortmann, 2008). On absolute scales too, low skill participants may be prone to overestimate just because they *perform* more poorly, having relatively more errors to mistake for successes. This potential bias, which we will call the 'performance artefact', would cause regressive estimates, but is conceptually and practically distinct from the well-mapped, and relatively easily controlled artefact of regression to the mean. Our path analysis for Experiment 1 suggests that this performance artefact may substantially govern the DKE.

These considerations highlight a broader confound, intrinsic to the DKE, yet which has received remarkably little explicit consideration. The effect is widely understood as a relationship between task skill and estimation error, but the measurement of estimation error has only ever been done in situations in which high and low skill participants differ in their success at the task. This confounds the broader concept of task skill (ability for a class of task) with a narrow measure of task performance (level of success at current instance of task), so does not 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 plainer 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 (i.e. success 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 ($\rho = -.30$), baseline performance and SET ($\rho = -.26$), and baseline performance and JND ($\rho = -.37$). Our methods and predictions were preregistered on the Open Science Framework <https://osf.io/ccgsz/> (see *Supplemental materials S1*). Experiment 2 was approved by the Psychology Research Committee, University of Edinburgh, as an extension to Experiment 1 (approval#105-16171). 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.⁶ 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 $\rho = -.30$ at $n=88$), and with JND (power .98 for $\rho = -.37$ at $n=88$), and adequate power to replicate the correlation with SET (power .80 for $\rho = -.26$ 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,

⁶ 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.⁷

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 ($\sim 50\%$).

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

⁷One 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).⁸ 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 ($p > 0.05$). In these cases, the psychophysical measures SET and JND could not be meaningfully estimated; five participants were excluded by these criteria.

⁸Dot 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 5 shows the successful titration to ~50% hit rate in the main block. The bottom two panels of Figure 5 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 5 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 6 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, the finding implies that the metacognitive differences are *not sufficient* for the DKE, in the absence of the usual performance differences between high and low skill participants.

Path analysis⁹

This conclusion can be further explored by a path analysis, structurally similar to that for Experiment 1, with three indirect paths mediating between task skill and estimation error: one path via main block performance; and dual metacognitive paths via SET and JND. The critical difference in Experiment 2 is that the strength of the links in the performance path have been experimentally pushed to zero, by eliminating systematic between-participant variance in main block performance. The path analysis is shown in Figure 7. The observed

⁹This path analysis was not part of our pre-registered plan, but was suggested by an anonymous reviewer.

performance path is null, as expected; but the metacognitive paths are strikingly similar to those for the Movement task in Experiment 1 (Figure 4a). Lax metacognitive calibration (high SET) tends to promote the DKE, but poor metacognitive sensitivity (high JND) attenuates the pattern, and the total metacognitive path is weak and not significant ($-.05$, $p = .67$). In the absence of performance differences, the DKE depends on metacognitive factors alone, and disappears ($-.05$, $p = .65$).

Experiment 2: 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 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 eliminated by the removal of performance differences in the main block. This supports the suggestion from Experiment 1, that a performance artefact is the main driver of the DKE, under typical circumstances in which high and low skill participants differ in level of success at a task. By contrast, metacognitive calibration and sensitivity had directionally opposite influences on estimation error, which largely cancelled out, so that metacognitive factors did not significantly mediate the DKE overall.

The above pattern describes the present data, but this does not mean that the dual-burden account could never provide an appropriate explanation of the DKE. Metacognitive and performance differences may conspire to create the DKE, as proposed originally by Kruger and Dunning (1999), in some tasks or circumstances. For instance, if the link between task skill and metacognitive calibration were very much stronger than in our data, then a substantial mediational role for metacognition could be envisaged. However, the dual burden account is not generally necessary to explain the DKE, since performance differences alone are capable of inducing the effect, assuming only that metacognitive insight is imperfect.

The key interactions can be captured by a simple graphical model, which shows how estimation errors would vary across two levels of metacognitive skill, and three levels of task performance, given a Gaussian distribution of response errors (Figure 8a).¹⁰ The headline outcome in each panel is the overall estimation error (EE). The dual-burden account assumes that performance and metacognition are tightly yoked, so that the DKE is specific to the contrast between a low performer with poor metacognition (bottom right panel, EE = 37), and a high performer with good metacognition (top-left panel, EE = 4). But, in fact, just as great a contrast would emerge between a low performer with poor metacognition (bottom right panel, EE = 37), and a high performer with equally poor metacognition (top right panel, EE =

¹⁰This illustration is a sub-set of the possible combinations of performance level, metacognitive sensitivity and metacognition calibration. A wider range of possibilities, more fully mapping the complex interactions between these factors, is provided in *Supplemental materials S3*.

4). Figure 8b further shows that a similar pattern can arise from performance differences alone, when participants have no metacognitive insight at all: compare the low performer with no metacognition (bottom panel, $EE = 47$) and the high performer with no metacognition (top panel, $EE = -4$). This last contrast is essentially Burson and colleagues' (2006) noise plus bias model, applied to absolute self-estimation. It is included to make the wider point that the DKE pattern alone does not imply metacognitive insight for anyone, let alone systematic metacognitive differences between skilled and unskilled participants.

Metacognitive differences may exist between high and low performers, as observed in Experiments 1 and 2, but such differences do not necessarily drive the DKE. Conversely, the DKE pattern does not imply underlying metacognitive differences, since the same superficial relationship between task skill and estimation errors can be driven by a performance artefact, provided only that participants have imperfect insight into their performance. In general, the metacognitive differences that are widely held to underpin the DKE are not established by the DKE itself, but must be demonstrated independently for any given task.

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; Krajč & 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 between competing accounts, showing that each of these factors can contribute to shaping the typical DKE. Our studies employ novel tasks, and online self-estimation. In drawing our main conclusions, we assume 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 showed good metacognitive abilities, being more sensitive, and having stricter standards, better calibrated to the objective criterion. These metacognitive differences correspond well with Kruger and Dunning's (1999) original concept of a metacognitive deficit amongst the unskilled. However, the effect size that these differences could account for was vanishingly small by comparison with the DKE as widely depicted (e.g. by the famous Figure 1 of Kruger & Dunning, 1999). This is because several other biases 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 $\sim .3$ for the Movement task and $\sim .2$ 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. However, a path analysis showed that this correlation was principally mediated by performance differences between high and low skill participants, rather than by differences in metacognitive insight. Low performers may overestimate more, just because they have more failures about which to be mistaken. This performance artefact is importantly distinct from the well-mapped influence of regression to the mean. The performance artefact is one part of the original dual-burden account of the DKE, but there has nonetheless been remarkably little discussion of the causal role of performance *per se*, as opposed to metacognitive factors. This may be because the dual-burden account assumes that performance and metacognitive factors are necessarily yoked; but the present study has shown that these influences are separable, both in principle and in practice. 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 are real, at least for our tasks, they are not sufficient for the DKE.

Moreover, considering that several other biases can induce the DKE, we must conclude that neither metacognitive differences, nor even the full dual-burden, are necessary for the effect. Estimation error depends in complex ways upon the particular mix of metacognitive sensitivity and calibration, and the distribution of response errors, and it 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 complex, the size of the response error more ambiguous, the objective criterion for success more opaque, and the feedback more cryptic. Furthermore, if making global self-estimates, rather than trial-by-trial judgements, uncertainty can only increase, especially if required to rank oneself relative to unknown others. We would suggest that metacognitive insight is highly unlikely to be a *more* direct determinant of estimation error in such complex scenarios. Of course, the role of metacognition in the DKE for any specific task or domain is an empirical question, and an important step will be to apply our analytical approach to the higher-level intellectual skills more typical of the DKE literature.

But if the broader scientific aim is to study metacognition, then researchers should strive to use measures that are free from the many confounds that estimation errors entail. Our adapted psychophysical approach is one possibility, appropriate given the distribution of response errors in our tasks. A range of more evolved alternative methods also exist, which can be applied where suitable tasks can be devised (Fleming & Lau, 2014).

None of our findings cast doubt on the DKE as an empirical phenomenon. On the contrary, our data extend the classic 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 misrepresentation 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 greatly 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 are thus as apt as ever. By contrast, the modern meme that *stupid people are too stupid to know they are stupid* is a dramatic oversimplification, 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.
- Augustyn, J. S., & Rosenbaum, D. A. (2005). Metacognitive control of action: Preparation for aiming reflects knowledge of Fitts's law. *Psychonomic Bulletin and Review*, 12(5), 911–916.
- 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.
- Campbell, D., & Kenny, D. (1999). *A primer on regression artifacts*. Guilford Publications.
- 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.
- Dunning, D., & Perretta, S. (2002). Automaticity and eyewitness accuracy: A 10- to 12-second rule for distinguishing accurate from inaccurate positive identifications. *Journal of Applied Psychology*, 87(5), 951–962.
- Edwards, R. K., Kellner, K. R., Siström, C. L., & Magyari, E. J. (2003). Medical student self-assessment of performance on an obstetrics and gynecology clerkship. *American Journal of Obstetrics and Gynecology*, 188(4), 1078–1082.
- 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.
- Kelley, C. M. ., & Lindsay, D. S. (1993). Remembering mistaken for knowing: Ease of retrieval as a basis for confidence in answers to general knowledge questions. *Journal of Memory and Language*, 32(1), 1–24.
- 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.
- Krajč, M., & Ortmann, A. (2008). Are the unskilled really that unaware? An alternative explanation. *Journal of Economic Psychology*, 29(5), 724–738.
- 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.
- Mabe, P. A., & West, S. G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology*, 67(3), 280–296.
- 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).

Parker, R. W., Alford, C., & Passmore, C. (2004). Can family medicine residents predict their performance on the in-training examination? *Family Medicine*, 36(10), 705–709.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36.

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.

Simons, D. J. (2013). Unskilled and optimistic: Overconfident predictions despite calibrated knowledge of relative skill. *Psychonomic Bulletin and Review*, 20(3), 601–607.

Tenenberg, J., & Murphy, L. (2005). Knowing what I know: An investigation of undergraduate knowledge and self-knowledge of data structures. *Computer Science Education*, 15(731797551), 297–315.

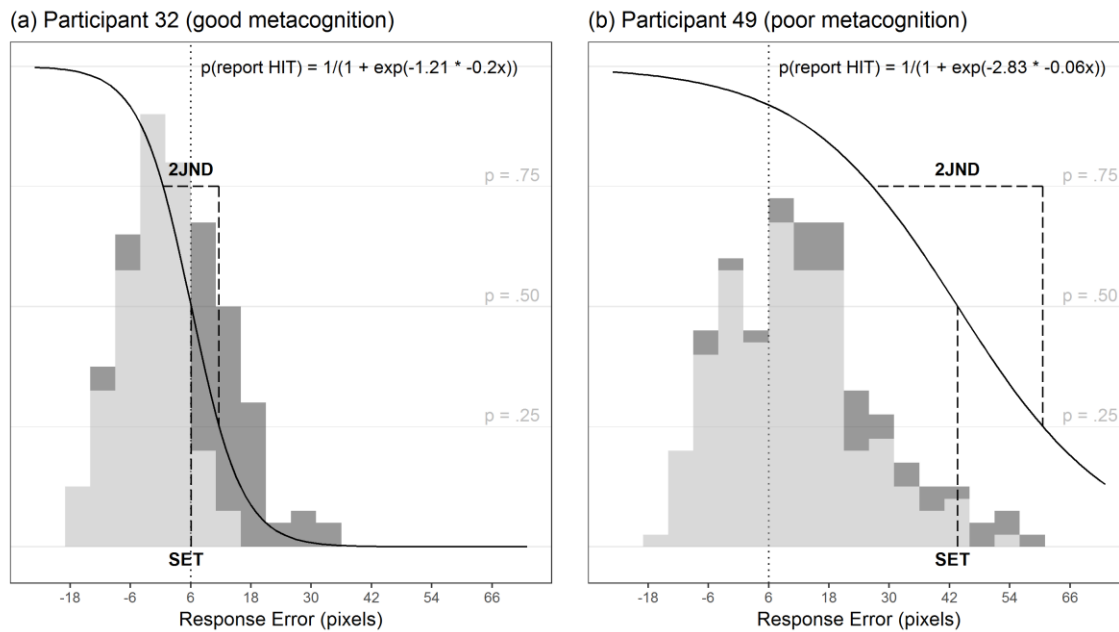


Figure 1. Examples of the analysis of metacognitive insight for the Movement task, for one participant with good metacognitive insight **(a)**, and one with poor insight **(b)**. In each panel, the histogram shows response frequency by size of response error, across 200 pointing movements, with negative 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 histogram is shaded by the frequency with which the participant reported a hit (light grey) or a miss (dark grey). The black curve is the logistic function relating the probability (p) of reporting a hit to the size of the response error, with the best-fitting equation shown at the top of the panel. Metacognitive calibration is measured by the subjective error threshold (SET), at which the participant is equally likely to report a hit or a miss, given by the x intercept for $p = .50$. Metacognitive sensitivity is measured by the just noticeable difference (JND), given by (half of) the difference in response error between $p = .25$ and $p = .75$.

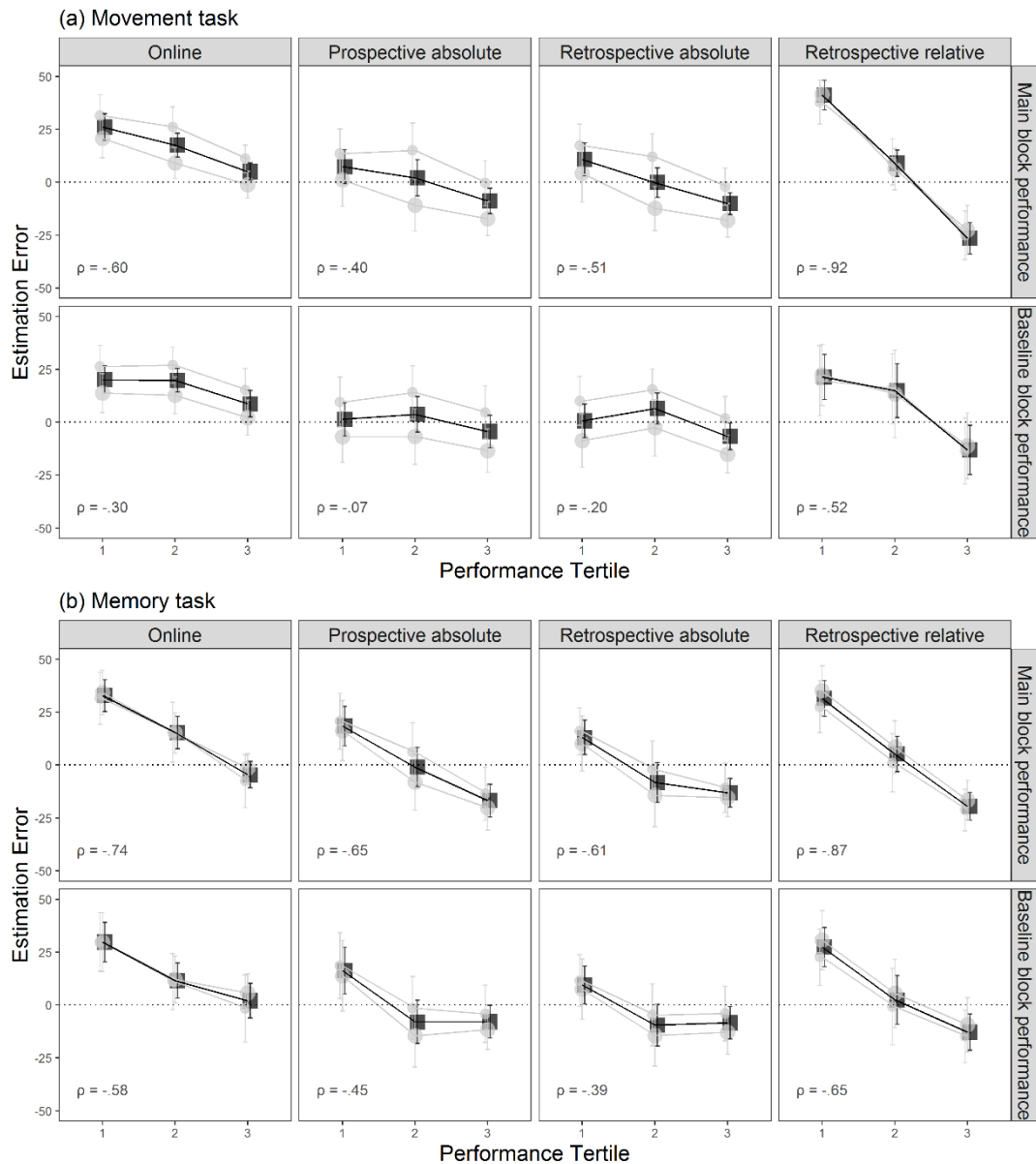


Figure 2. Experiment 1. Relation between performance and estimation error for **(a)** the movement task ($n = 80$) and **(b)** the memory task ($n = 62$). Performance is defined by hit rate in the main block (upper row) or the baseline block (lower row). Estimation error is derived from online self-estimation, 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 (\pm between-subject 95% CIs) across all targets. Small and large grey circles show means (\pm 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.

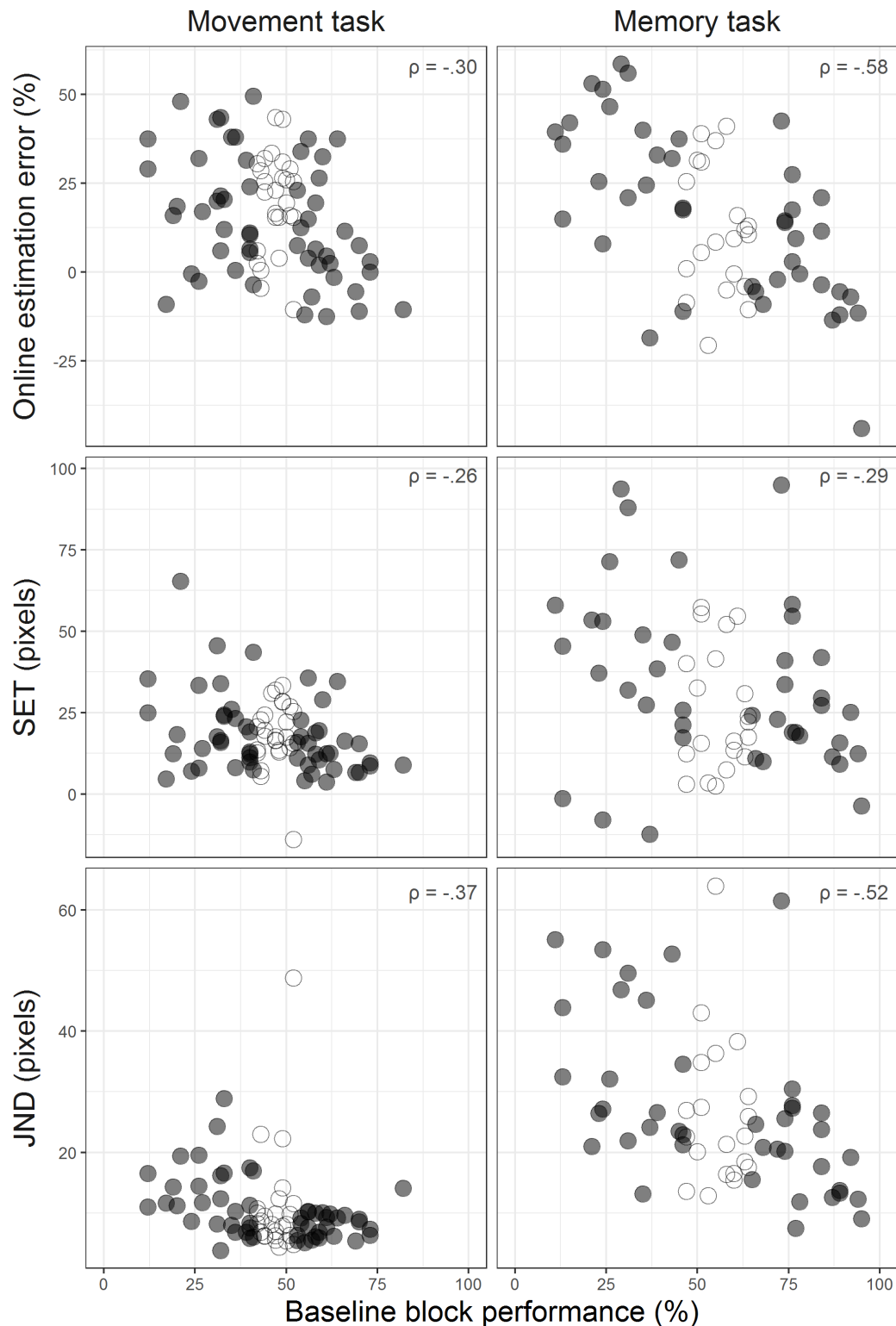


Figure 3. Experiment 1. Relation of baseline block performance (% hit rate) to online self-estimation measures for the main block, for the movement task ($n = 80$) and memory task ($n = 62$), with Spearman's ρ for each plot. Participants in the middle tertile of performance are plotted as unfilled dots 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).

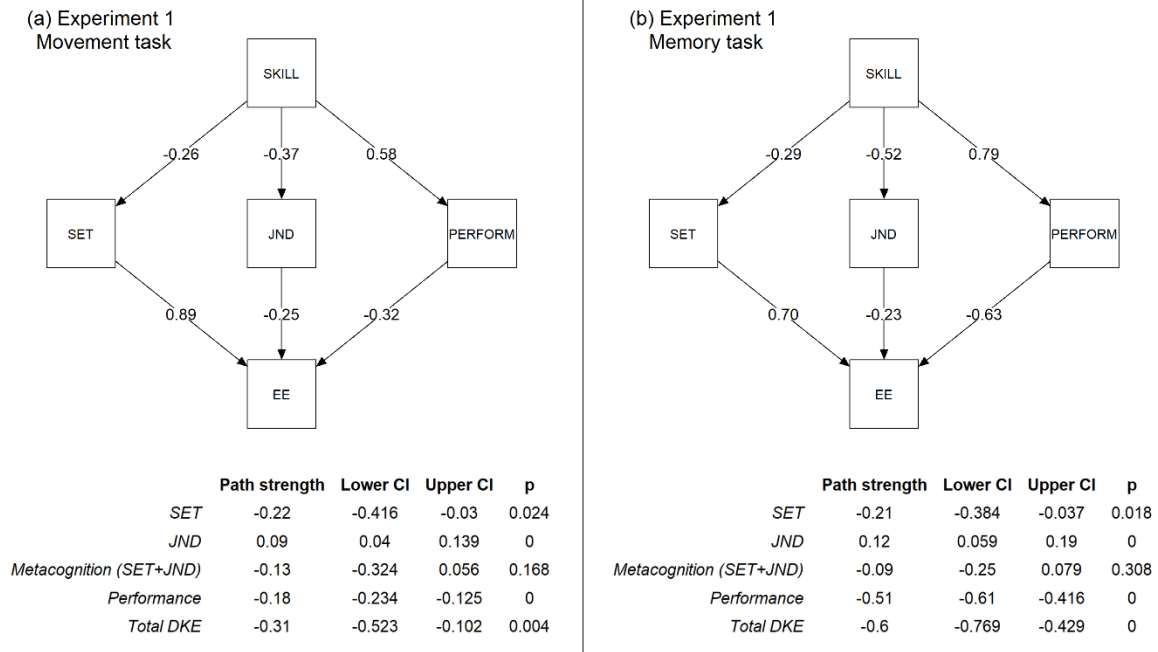


Figure 4. Path analysis for Experiment 1 **(a)** Movement task and **(b)** Memory task. The DKE is represented by the relationship between task skill (SKILL = hit rate in baseline block) and online estimation error (EE in main block). The variables hypothesised to mediate the DKE are metacognitive calibration (SET), metacognitive sensitivity (JND) and task performance (PERFORM = hit rate in main block). The standardised path strength (correlation coefficient) is shown on the diagram for each link. The estimated strength of each path (given by the product of its two links) is tabulated, with 95% confidence intervals and associated p value. The summed path strengths for SET and JND give an estimate of the total mediational effect of Metacognition. The summed path strengths for Metacognition and Performance give the total DKE. Also included in the model, but omitted from the figure, are the correlations between SET, JND, and PERFORM, to ensure that the null hypothesis of a well-fitting model is not rejected (Movement task $\chi^2 = 1.08$, $df = 1$, $p = .30$; Memory task $\chi^2 = 2.01$, $df = 1$, $p = .16$).

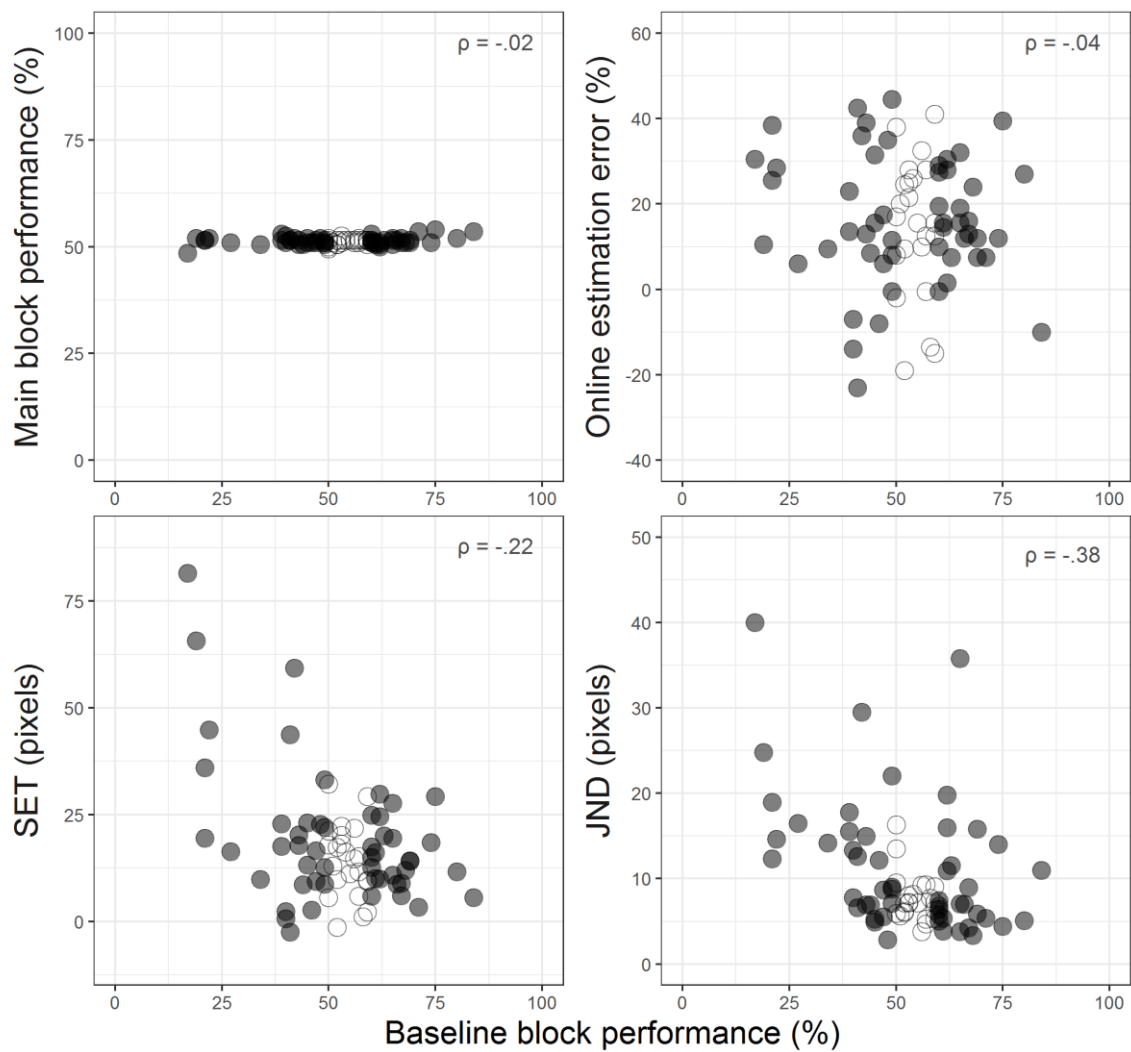


Figure 5. Experiment 2 ($n=75$). Relation of baseline block performance to (levelled) main block performance, and three online self-estimation measures, with Spearman's ρ for each plot. Participants in the middle tertile of task skill are plotted as unfilled dots to visually separate performance tertiles. 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.

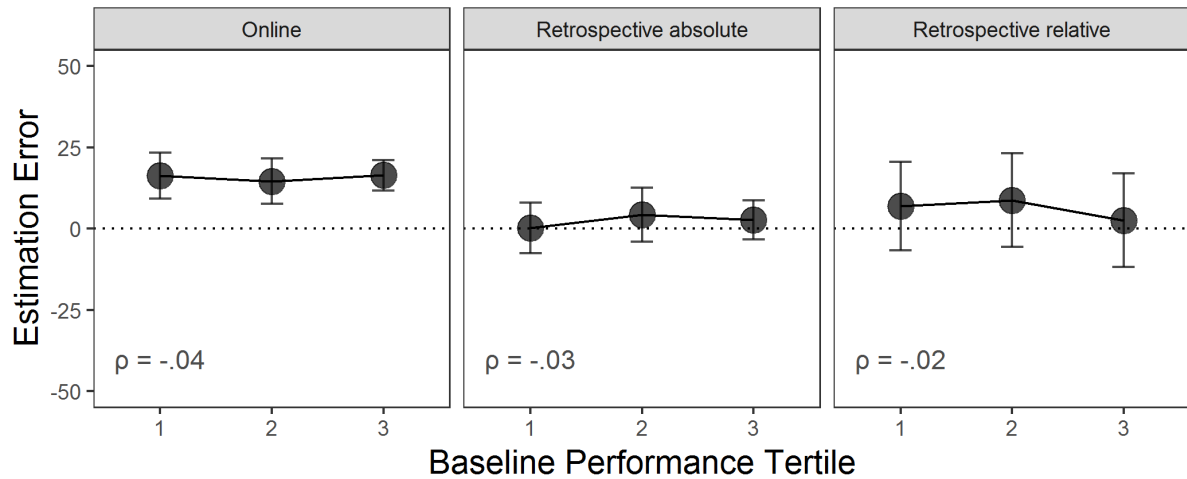


Figure 6. Experiment 2 ($n = 75$). Relation between baseline performance and estimation error. Estimation error is derived from online self-estimation, a retrospective absolute estimate, or a retrospective relative (percentile) estimate. The means are split by baseline performance tertile (where 1 is 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 meaningless in this context. 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).

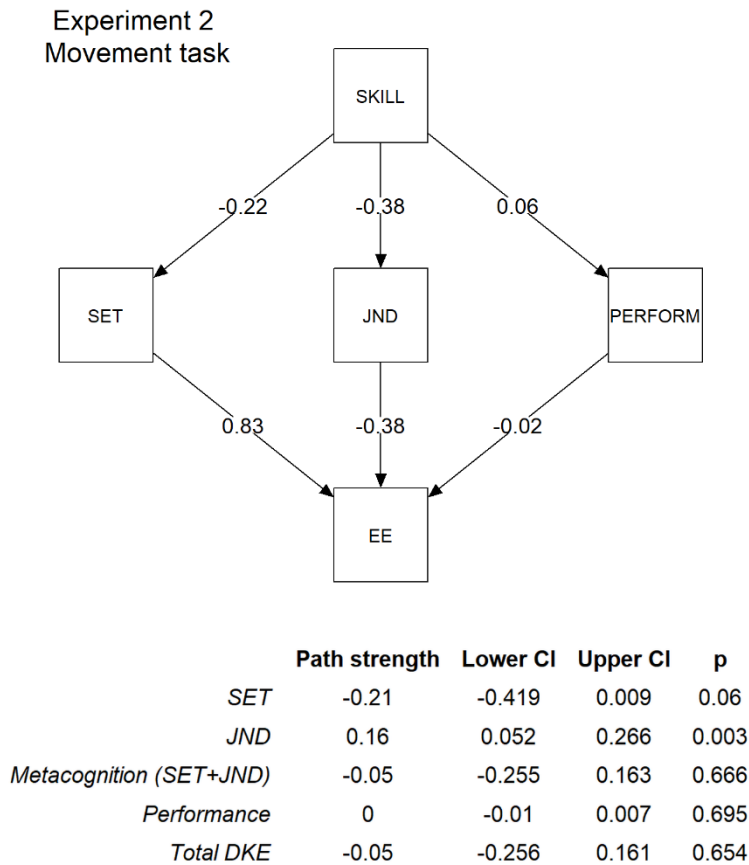


Figure 7. Path analysis for Experiment 2, similar to that for Experiment 1 (Figure 4). The DKE is represented by the relationship between task skill (SKILL = hit rate in baseline block) and online estimation error (EE in main block). The variables hypothesised to mediate the DKE are metacognitive calibration (SET), metacognitive sensitivity (JND) and task performance (PERFORM = hit rate in main block). The standardised path strength (correlation coefficient) is shown on the diagram for each link. The estimated strength of each path (given by the product of its two links) is tabulated, with 95% confidence intervals and associated p value. The summed path strengths for SET and JND give an estimate of the total mediational effect of Metacognition. The summed path strengths for Metacognition and Performance give the total DKE. Note that the links in the performance path have been experimentally pushed to zero by our performance levelling technique, so the DKE is wholly mediated by Metacognition, and is effectively absent. Also included in the model, but omitted from the figure, are the correlations between SET, JND, and PERFORM, to ensure that the null hypothesis of a well-fitting model is not rejected ($\chi^2 = 0.02$, $df = 1$, $p = .89$).

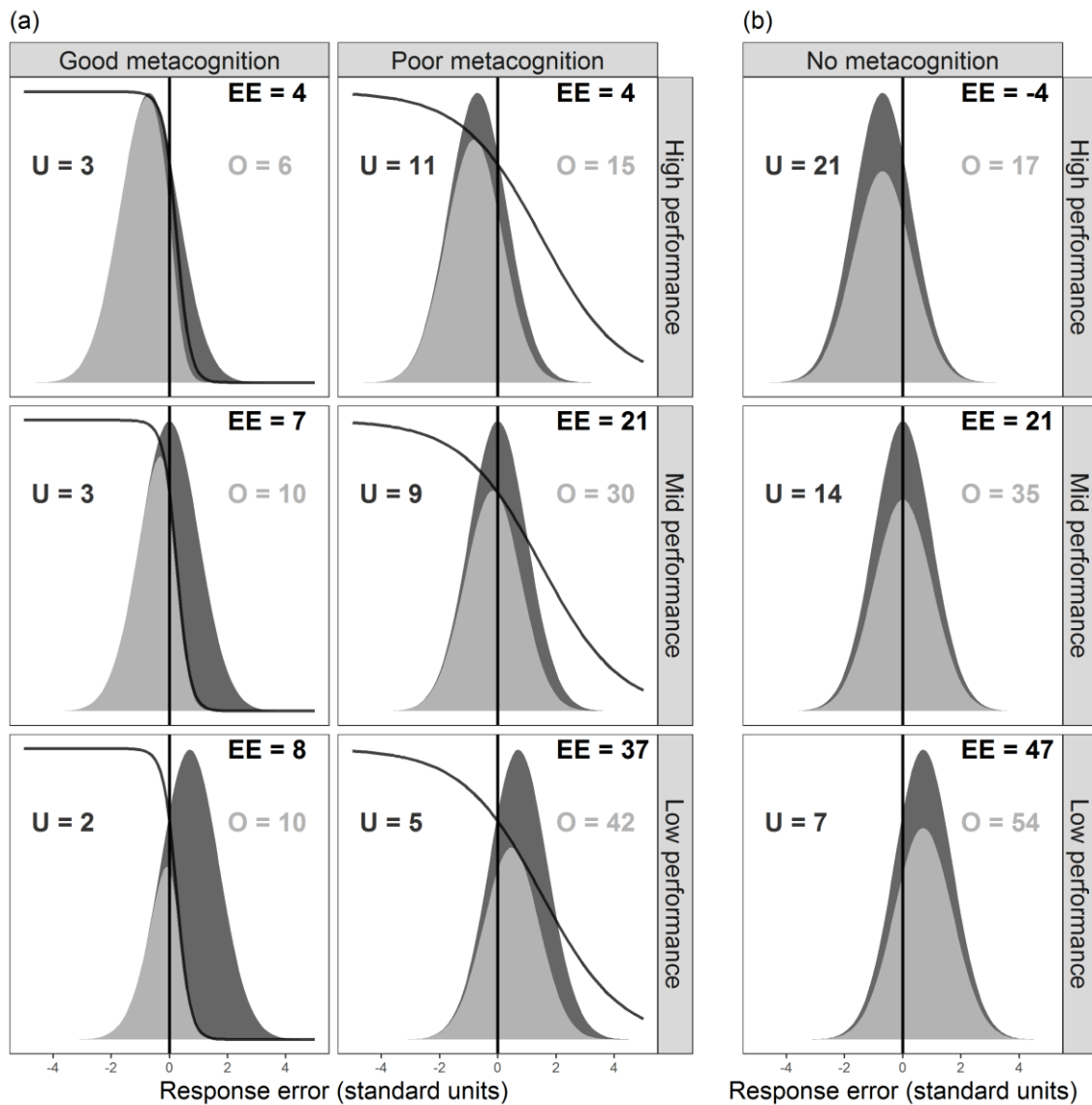


Figure 8. (a) Illustration of the combined effects of metacognitive insight and performance level (i.e. success rate) on total estimation error. Each plot shows a normal distribution of errors, shaded according to the proportion of hit reports (light grey) and miss reports (dark grey) at each level of error, tracked by a psychophysical function (curved black line). Good metacognition has a steep function, centred relatively close to the objective threshold for success (vertical black line). Poor metacognition has a shallower function, centred further rightward. Underestimation occurs when a miss is reported for an objective hit, so the total underestimation (U) is given by the dark grey area to the left of the vertical black line. Overestimation occurs when a hit is reported for an objective miss, so the total overestimation (O) is given by the light grey area to the right of the vertical black line. Estimation error (EE) is given by the difference (O-U), with all values expressed as a percentage (to the nearest percent) of total responses. EE varies with metacognitive insight, but it also varies with level of performance on the current instance of the task. These influences interact. **(b)** Participants are here modelled to have no metacognitive insight at all, but to show an (optimistic) response bias, reporting hits on a majority of trials regardless of response error. This can also produce the pattern of gross overestimation amongst low performers, and more accurate estimation amongst high performers.