Overconfidence, Preference for Control, or Unskilled and Unaware?*

Jean-Pierre Benoît
London Business School
jpbenoit@london.edu

Juan Dubra
Universidad de Montevideo
dubraj@um.edu.uy

Giorgia Romagnoli
University of Amsterdam
G.Romagnoli@uva.nl

June, 2018.

Abstract

We report the results of five experiments where we address the nature and measurement of Overconfidence. First, we develop model to define and evaluate various possible definition of control, and estimate each one. In second term, we develop a new method to measure beliefs that accounts for the tendency of people to bet on devices where their skill might be involved (a "bias" called Control in the psychology literature). This allows us to improve over the methods traditionally used to measure beliefs' about one's own performance. The influence of Control in the measure of overconfidence in this second experiment is significant and the point estimate is 6%: true average belief of being in the top half in a test is 59%, but when we do account for Control, individuals report 65%.

In three other experiments we address the hypothesis of Kruger and Dunning, "unskilled and unaware". Their hypothesis is that populations exhibit overconfidence because the unskilled are unaware of their lack of skill (eg: the same skills needed to

*We thank Facundo Danza for outstanding research assistance.

write a grammatically correct sentence are those needed to know that the sentence is correct). We test this hypothesis against the null that the whole population (not just the unskilled) is overconfident, but regression to the mean makes the unskilled look overconfident (eg: the beliefs of the unskilled are a weighted average of the prior –"I am average"– with the signal –"I did poorly in this test"). Using a variety of tests (on the experiment in Benoît, Dubra and Moore, and in a new experiment) we cannot reject that the whole population is overconfident. To the best of our knowledge, this is the first statistical test of the "unskilled and unaware" hypothesis.

Finally, on a fifth experiment we explore the same issue of unskilled and unaware vs general overconfidence, studying directly the (structurally estimated) signalling structures elicited in the experiment. This also allows us to verify whether each individual is also unaware when he is unskilled, vis a vis those questions where he is skilled. Our estimations show that one cannot reject that the signalling structures of the unskilled are as good as those of the skilled.

VERY PRELIMINARY. DO NOT CITE.

Keywords: Overconfidence, Control, Unskilled and Unaware. Experimental Methods.

Journal of Economic Literature Classification Numbers: D3

The standard method for eliciting beliefs about an event is to offer a series of choices between bets of the sort "do you prefer a bet that pays \$10 if the event E happens, or a bet that pays \$10 with probability $x\%$", where x varies from 0% to 100%. If the person thinks the event has a probability 34% of happening, he will choose to bet on the event for $x < 34$, and to choose the random bet for $x > 34\%$ (see Benoît, Dubra and Moore, Eil and Rao (2011) and Merkle and Weber (2011) and Burks et al. (2013)). In the study of overconfidence, and the better than average effect, the event E is of the form "your test score is in the top half of scores", and rationality is rejected if the average of all reported probabilities is larger than 50%.

The problem with this methodology is that even if my belief is (say) 34% that my score will be in the top half of test takers, I might still bet on my test score being in the top half, if the choice is against a bet that pays with probability 40%, because I like to bet on things where my skill is involved. In psychology, this bias (which is different from overconfidence, as my belief is still 34%) has been called Control (see Heath and Tversky (1991), Goodie (2003) and Goodie and Young (2007)). The basic hypotheis is that, even if people cannot

affect the outcome of the bet on self, they might still prefer to bet on a device where, with enough practice or training, they might be able to modify the probability of winning.

Therefore, even if some papers have found overconfidence (BDM, Eil and Rao, Merkle and Weber, or Burks et al.), the correct inference is only that a model of rationality *and* no control is rejected. If the desire for control is large enough, beliefs might be correct, and measured overconfidence might be the result of control. Our first contribution is to develop a method of measuring beliefs that is not affected by control. Our (preliminary) findings indicate that on average people report a probability of 67% of being in the top half in a task (under rationality and no control it should be 50%), and that the effect of control is significant but small (around 3-5%).

Our second contribution is to develop two methods to ascertain whether the Kruger-Dunning hypothesis that populations are overconfident because the unskilled are unaware is correct. Since we observe that populations display overconfidence, with high types appearing slighty underconfident while low types appearing very overconfident, two competing explanations have arisen. According to Kruger and Dunning, the main reason for measured overconfidence is that the unskilled lack the knowledge to realise they are unskilled, and therefore are unable to evaluate that their performance is low (because the skills needed to perform well are the same as those needed to evaluate performance). A stylized version of that theory is therefore that the skilled are well calibrated, while the unskilled are overconfident and are less able to recognize signs that they performed poorly.[1] The other popular alternative explanation is that everybody is overconfident (not just the unskilled) but that there is regression to the mean: unless signals are perfectly informative, beliefs about the probability of placing in the top half in some task will be an average of the prior and actual placement, so for someone who performed in the bottom half, their belief will necessarily be overconfident. In their influential work (4.000 + cites), and in follow up papers, Kruger, Dunning and co-authors have tried to argue that the right explanation is that the unskilled are unaware. However, to the best of our knowledge, there has been no formal test of the theory in the form of a statistical test.

We develop two statistical tests that show that one cannot reject the model of "(generalized) overconfidence + regression to the mean" in favor of "unskilled and unaware".

[1]Strictly speaking, although less central in their original paper, the skilled might be underconfident because they fail to realize how good they are; what looks easy to them must be easy for everybody.

So far, we have run five experiments. Two where we measure control, and three where we compare "unskilled and unaware" with "generalized overconfidence + regression to the mean". We first address control, and then unskilled and unaware.

1 Control

We present the results of two experiments, and a structural model to evaluate one of them.

1.1 Experiment 1 and structural model.

In this section we provide a model of preferences that have a taste for control in "all" of its forms. The result is a two parameter family of preferences that allows us to pin down behavior in each treatment. Concretely, we show that the elicitation mechanism is incentive compatible (when control is not present) and we can find the optimal response of subjects (to the choice of a "declared belief") in each treatment, as a function of both parameters.

To understand the model, and how the different versions of control are modeled, we now describe the versions in the particular context of this experiment.

We will model an individual who has exerted some effort in order for something to come true: the subject wants his or her score to be high, lets say in the top half. Lets call E the event "your score in the top half".

We will focus on bets of three kinds, all with prizes of \$0 or \$10: bets that have objective randomness, bets that pay if the "favorable" event E happened, and bets that pay where the complement of E happens (E^C).

In the experiment, in treatment 1, there is a bet that pays \$10 if E happens, and the individual must choose a probability p with which that bet is played; with probability $1-p$ a bet that pays \$10 with probability $\frac{1+p}{2}$ is played.[2] In principle, the belief elicited this way could be contaminated if the individual likes to play bets where his skill is involved.

In treatment 2 there is a bet that pays \$10 if E^C happens, and the individual must choose a probability p with which that bet is played; with probability $1-p$ a bet that pays \$10 with probability $\frac{1+p}{2}$ is played. Here, the elicitation of beliefs could be contaminated if the

[2] A random number $\omega \sim U[0,1]$ is chosen. If $\omega \leq p$, the individual earns \$10 if E happens and 0 otherwise. If $\omega > p$, the individual earns \$10 with probability ω. Since $E(\omega \mid \omega > p) = \frac{1+p}{2}$, we say that if $\omega > p$, the lottery pays with probability $(1+p)/2$.

individual likes to play bets where his skill is involved, even if it is involved in a negative way (him "losing" in terms of the score being in the bottom half).

In treatment 3 there is a bet that pays \$10 if E happens, and a bet that pays \$10 if E^C happens, and the individual must choose a probability p such that with probability p the E bet is played, and with probability $1-p$ the E^C bet is played. [3] Here the mechanism would not distort the belief if the individual likes to bet on himself, regardless of whether the event involved is good or bad, but would inflate or deflate the belief if the individual likes to bet on nice events or bet on negative events.

For w the initial wealth, and u an expected utility over wealth levels (with $u(w+10)=1$ and $u(w)=0$), and μ the subjective probability that the "winning event" will happen (E in treatment 1, and E^C in treatment 2), in the traditional setting without control the individual chooses p to maximize

$$\begin{aligned} U(p) &= p\left[\mu u(w+10)+(1-\mu)u(w)\right]+(1-p)\left[\frac{1+p}{2}u(w+10)+\left(1-\frac{1+p}{2}\right)u(w)\right] \\ &= p\mu+(1-p)\frac{1+p}{2} \end{aligned}$$

which is of course maximized for $p=\mu$.

Suppose that the subject can choose the probabilities with which he can play a random bet (call the probability of playing such a bet x), a "favorable bet" where he wins if he did well (call the probability y) and a negative bet where he wins if he did badly (call the probability z). We will add three terms to the traditional utility: $0.x$ (the subject doesn't care about betting on a random bet, other than for the material utility), $m.y$ and $n.z$. Those terms represent the utility the subject has from placing bets in his favor, or against himself. In principle m is positive, and n could be positive (if control means the subject likes to bet on things where his performance is involved) or negative (if he doesn't like to bet against himself) or 0.

Suppose μ is the belief in the "paying" event, and y and z are the probabilities of the favorable and unfavorable bets; in the general case where the three bets are allowed, and

[3] An alternative would be: the subject chooses a p and two numbers are drawn; if the first is below p, the E bet is executed, and if the second is below $1-p$ the E^C bet is executed.This is not the same as choosing one number p, and only drawing one random number such that if it is lower than p E is executed, and otherwise E^C is executed. In that case the individual would want to choose p either 0 or 1, depending on whether E or E^C is more likely.

It is easy to show that this does not elicit the belief, even in the absence of "control" concerns.

when all three forms of control are allowed the utility is

$$U(p, y, z) = p\mu + (1-p)\frac{1+p}{2} + my + nz. \tag{1}$$

In treatments 1 and 2 the utility becomes

$$\begin{aligned} U_1(p) &= p\mu_E + (1-p)\frac{1+p}{2} + mp \Rightarrow p^* = \mu_E + m \\ U_2(p) &= p\mu_{E^C} + (1-p)\frac{1+p}{2} + np. \end{aligned} \tag{2}$$

In the latter case the elicited p would be the probability of E^C, but if we keep p for the probability that E happens, the alternative way to write the utility is

$$U_2(p) = (1-p)(1-\mu_E) + p\frac{2-p}{2} + n(1-p) \Rightarrow p^* = \mu_E - n$$

and then the elicited p is still that of E.

In treatment 3, the individual chooses a p, and a coin is flipped. If heads comes up, a random number is selected and the individual is paid if E happens when the number is less than p, or paid with probability $\frac{1+p}{2}$ if the number is above. If tails comes up, a random number is drawn, and the subject is paid if E^C when the number is below $1-p$, or paid with probability $\frac{2-p}{2}$ if the number is larger than $1-p$. In order to specify the utility in equation (1) for this case, we assume that the utilities my and nz are enjoyed even if the bet is not played (like paying the cost of effort, even if it turns out to not be needed; or the utility of saying "if a flood comes, I will donate money", and then the flood doesn't come you still enjoy the utility). In this alternative formulation, the sizes of m and n "double" in importance relative to the alternative forumlation in which the utility is only obtained if the relevant bet is realized. Concretely, we have

$$U_3(p) = \frac{1}{2}\left(p\mu_E + (1-p)\frac{1+p}{2}\right) + \frac{1}{2}\left((1-p)(1-\mu_E) + p\frac{2-p}{2}\right) + mp + n(1-p). \tag{3}$$

That is, the mp and $n(1-p)$ are not affected by the $\frac{1}{2}$ (in the alternative formulation they are inside the corresponding bracket). This is an important assumption, as identification depends on it. Whether the assumption is valid, or not, is a psychological matter that we have not seen discussed. The optimal p is given by

$$\frac{dU_3(p)}{dp} = m - n - p + \mu_E \Rightarrow p^* = \mu_E + m - n. \tag{4}$$

We are now in a position to define formally the three forms of control:

- strong control is when the individual likes to bet in any bet where his skill is involved, so that $m, m > 0$ (possibly, $m \geq n$).
- weak control is when the individual likes to bet on bets that pay when he did well, and is indifferent to negative bets, so that $m > 0 = n$.
- positive control is when he likes favorable bets, and dislikes negative bets, so that $m > 0 > n$.

This establishes that with strong control, with $m = n$, there is no distortion in treatment 3. Also, if there is weak control, the upward bias is the same in treatments 1 (in equation 2) and 3 (equation 4).

In the positive control case ("you don't like to bet against yourself" just as strongly as you like to bet on your success), we set $n = -m$. In that case the bias is twice that of treatment 1: the subject has a utility for betting on himself of mp, and one of not betting against himself of $m(1-p)$, so it is counted "twice".

This setup will allow us to see whether the treatments permit a distinction between overconfidence, and estimate all versions of control.

1.1.1 On identification

Let μ_E denote the subjective true probability of being in the top half. μ_E may include overconfidence and is the object we want to estimate. Let's call p_1 the average reported probability in treatment 1 (where we pay in the event subjects are in the top half), p_2 the average reported probability (of being in the top half) in treatment 2 (where we pay in the event subjects are in the bottom half), and p_3 the average reported probability in treatment 3 (where we toss a coin to determine whether we pay according to the placement in the top or the bottom half). From section 1.1 we have:

$$\begin{aligned} p_1 &= \mu_E + m \\ p_2 &= \mu_E - n \\ p_3 &= \mu_E + m - n \end{aligned}$$

Given the observed vector $\{p_1, p_2, p_3\}$, the above system can be solved for the unobservables and gives the following:

$$n = p_1 - p_3 \quad (5)$$

$$m = p_3 - p_2 \quad (6)$$

$$\mu_E = p_1 + p_2 - p_3 \quad (7)$$

This means that our design allows us to estimate:

- The true unbiased subjective probability of being in the top half μ_E. *Note:* μ_E will still incorporate overconfidence and the tendency to say nice things about oneself- (the latter we will be unable to estimate in these experiments).
- The bias induced by betting on control for favorable events m
- The bias induced by betting on control for unfavorable events n.

From the data and equations 5-7 we will also be able to check which of the three versions of control is present on average in the population.

To illustrate. Suppose people have a weak control bias, so that $m > 0 = n$. If $p_1 > \frac{1}{2}$, is that overconfidence ($\mu_E > \frac{1}{2}$) or weak control ($m > 0$)? To understand in words the identification mechanism, notice that the sum of the probabilities assigned to "being in the top half" (in treatment 1), plus the probability of "being in the bottom half" (in treatment 2), should equal 1; whatever the difference with 1 is can be assigned to control. In fact, from section 1.1,

$$p_1 + (1 - p_2) - 1 = \mu_E + m + (1 - \mu_E + n) - 1 = m. \quad (8)$$

The idea here is that if the individual is indifferent among a random bet and a bet against himself, whatever he says in treatment 2 about his chance of being in the bottom half is "correct" (i.e. $1 - p_2 = 1 - \mu_E$); suppose his belief of E^C is 40%, so that $1 - p_2 = 40\%$. If in treatment 1 the individual reports that his "belief" of placing in the top half is $p_1 = 70\%$, then we know that he is inflating the total probability by $m = 70 + 40 - 100 = 10\%$. That is the content of equation (8).

1.1.2 Data and Results.

xl file: Data Control BDM version

This experiment was run entirely with the undergraduate population at the University of Amsterdam.

Basically the 3 treatments exhibit the same average overconfidence: in treatment 1 (68 subjects) the average reported probability of being in the top half was 66.2%; for the second (61 subjects) it was 67.9, and for the third (67 subjects) 66.7. There are large standard deviations of comparable magnitude across treatments (16.9, 18.5 and 19.8 for treatments 1-3 respectively).

We performed two tests. With the Wilcoxon rank sum (Whitney-Newey) test the p value for equality of distributions for treatments 1 and 2 was 61% (we would reject with a p value lower than 10%, say). For treatments 2 and 3 it was 76%, and for treatments 1 and 3 it was 78%.

We also ran the corresponding t test for difference of means, and we do not reject equality (Treatments 1 and 2, p value of 58%; treatments 2 and 3, p value of 72%; treatments 1 and 3, p value of 88%).

These results indicate that the measurement of beliefs is robust to the control manipulations, and they seem to indicate that control is not an issue in the measurement of overconfidence. While the estimated m, n and μ_E from equations 5-7 yield $m = -1.2$, $n = -0.5$ and $\mu_E = 67.4$, we cannot reject that $m = n = 0$.

1.2 Experiment 2, new design for control.

A different approach to the one pursued in the previous experiment is to remove the issue of control from the elicitation mechanism, instead of trying to pitch different versions of control against each other in the elicitation of various bets. The basic problem of the standard elicitation procedure is that the individual has to choose between a bet in which his skill is involved and a bet in which it isn't; if there is a desire for control, he will choose the bet where his skill is involved more often.

In this section, instead, we elicit beliefs by letting the individual choose between two bets where his or her skill are involved. To understand how we estimate control, we first describe a new version of the "standard" elicitation mechanism, against which the control elicitation mechanism will be compared. In both treatments, the subjects first play a visual game, in which a string of numbers appears blinking on the screen, and the individual then has to write it on a box (the dificulty varies in the amount of time the number appears, and in the

length of the string); call s the success rate of the individual out of the 10 repetitions of the task. Then, subjects see three sample questions which a similar to the ones they will have to answer in a Logic and Trivia Quiz. Then they are asked to report their belief that they will place in the top half of test takers. In the standard treatment, the elicitation procedure is similar to the one in Experiment 1, but the prize to be won is 10 lottery tickets, each worth a 3% chance of winning 20 Euro.[4] Individuals have to name a p; we draw a random number x; if $x \leq p$, the individual wins the lottery tickets if his score was in the top half of test takers; if $x > p$ the individual wins the lottery tickets with probability x. It is easy to check that this mechanism elicits the individual's belief.

In the Control treatment, the choice is not between betting on the individual's placement or on the random number x, but rather between the placement and the performance in the visual task (call s the success rate in the visual task). In particular, the individual names a p; if a random draw of x is $x \leq p$, the individual wins the lottery tickets if he placed in the top half in the quiz (as in the standard treatment); if $x > p$, the individual earns $10\frac{x}{s}$ lottery if she was sucessful in a randomly drawn round of the visual task, so if $x > p$ the probability of winning is $10x$, but the realization of the bet depends on the success in the visual task. That is, the mechanism is incentive compatible, and the randomness depends on the subjects's skill. The individual is not told that the number of tickets is $10x/s$, but rather that if he was successful in a randomly chosen round, he would earn **N** tickets (the **N** is varied according to the success rate of the subject in the task), and that it is in his best interest to report his true belief (see the appendix for the instructions). In this way, if the individual has no incentives to distort p away from his belief, even if he likes to bet on "devices" where his skill is involved (in one case his skill in the quiz is involved; in the other his skill in the visual task).

We pre-registerd the experiment (this one, plus experiments 4 and 5 on "unskilled and unaware"). Three hundred and thirteen subjects participated. In the standard treatment subjects report a 64.7%% chance of being in the top half on average, while they report 61.4% in the Control treatment. The 3.3% difference is significant at 7% if we don't depurate the estimate of the noise introduced by other factors (subject gender, etc.). **Giorgia, introduce here the result of the regression using controls** Interestingly, although the quiz was somewhat harder than the one we used in Benoît, Dubra and Moore (average of 7/12

[4] As is well known, paying in lottery tickets induces risk neutrality.

in this quiz, while 18/20 in BDM), and in Experiment 1, the average reported probability is very similar (67% in BDM, and in the three treatments in Experiment 1).

add somewhere that we tell them in any case where they placed, so that this does not inflate the estimate of overconfidence, or of belief

add also the discussion about the effect of control on the estimation (it is marginally significant), or the effect of control on the estimation of overconfidence (ceases to be significant). To me, just the fact that we quantify control is relevant. But you guys seem to think that we should re-pitch the whole thing to highlight that we care about control, over and above its effect on the estimation of overconfidence

1.3 Discussion

In Experiment 1 we varied the underlying phenomenon about which subjects bet (placing on top, placing on bottom, or a mix between both), and found no effect of this form of control. One interpretation is that in the elicitation mechanism subjects are indifferent among bets where the underlying event is varied from positive, to mixed, to negative.

Experiment 2 presents a more direct test of control in which we try to eliminate control concerns from the elicitation mechanism. We make the random device that determines success, when the subject does not bet on his skill in the test, depend on the subject's ability in a visual task. In this way, the reported belief of being in the top half falls by 3.3% relative to the standard elicitation mechanism (this is marginally significant, **we are currently including controls in a regression to increase the precision of this estimate**). That is, we show that control inflates the reported belief of being in the top half.

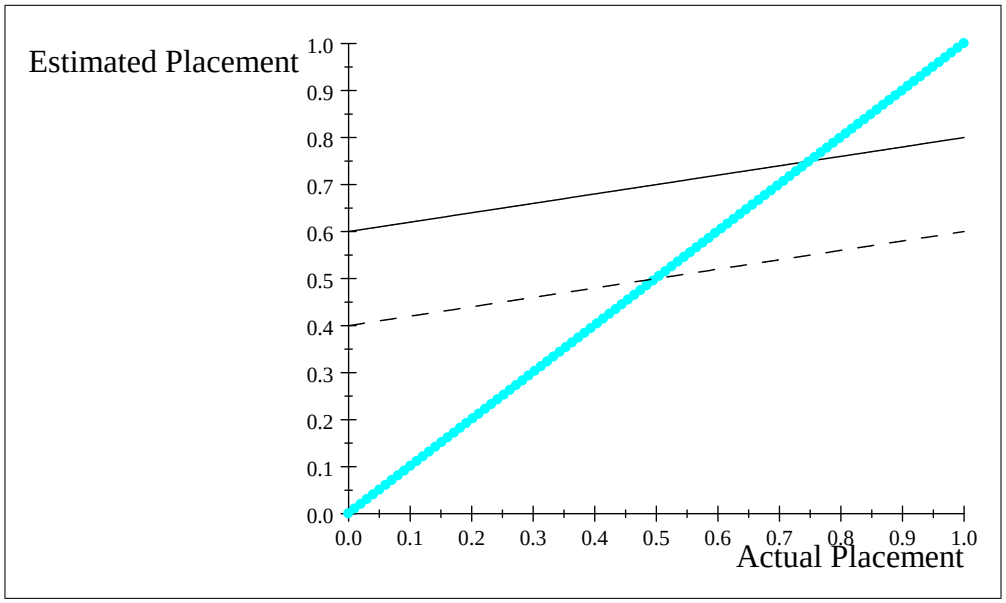
By increasing further the "importance" of the skill used to determine the chance of success in the elicitation mechanism (what was here the visual task), it might be possible to further sharpen the estimate of overconfidence. That is, if people care more about betting on their skill in the test, than betting on the visual task, our estimate of overconfidence might still be inflated by control.

2 Unskilled and Unaware.

In this section we report the results of three experiments that study whether the cause of measured overconfidence is that the unskilled are unaware, or that people in general are

overconfident and there is regression to the mean.

To understand why the two theories compete, consider the following "stylized fact": the skilled are mildly underconfident, while the unskilled very overconfident. For instance, if we plot actual percentile placement on the x axis, and estimated percentile placement in the y axis, a graphical representation of the stylized fact would compare the extimated percentile placement with a perfectly calibrated population (the 45° line) to obtain the following Figure.



We have also plotted, in dashed, the estimates of a population that is well calibrated on average, but where information is imperfect. In particular, a person who obtained the highest score must (a fortiori) be underconfident: her estimate will necessarily be weakly below her placement. The noisier the information, the flatter then dashed line.

Kruger and Dunning argue that the unskilled are overconfident because the skills needed to perform well are the same as those needed to evaluate how well you did on a test; since they lack the skills to perform well, they fail to realize that they performed badly too. They also argue that the skilled fail to realize how smart they are, so that they fail to take into account that the task is not as easy for others as it is for them. This joint hypothesis would explain the stylized fact. But it can also be explained by the nature of the noisy information (which would yield reports as in the dashed line), plus a "uniform" level of overconfidence, which would yield the actual estimated self placements.

2.1 Experiment 3, Unskilled and Unaware, Data from BDM.

Experiment 2 in BDM was not designed to study the hypothesis of Unskilled and Unaware, but the data can be used to study it, and in fact the paper contains a few comments about it. In that experiment, people were shown 5 sample questions, and although they were not incentivized, and their answers were recorded. Then they had to report what they thought was their chance of placing in the top half (they were also given information that the median score was 18/20).

In this section we use the data to estimate a structural model where the two hypothesis can be embedded, and we ask whether one can reject the model where all subjects are alike (same level of overconfidence and same signalling structure, which amounts to saying that skilled and unskilled process information in the same way) or the unskilled have a more noisy signalling structure which does not allow them to figure out that they are unskilled.

In the experiment, most people get 3, 4 or 5 (sample) questions right, so to simplify the model, we assume that they observed only 2 sample questions, and that the L, M, H signals would correspond approximately to the idea that they got 0,1 or 2 correct (we know their marks in the sample, but not their signals; they know their signals, not their marks). In the experiment they report their estimated chance of placing in the top half; given that the median was 18, we use their reports as their belief that they will score 19 or 20 in the quiz (the mapping is not exact, but it is inmaterial; this is just to develop the tests).

In principle, a "general model" is the following. There are two types, those who answer a question with probability q and those who do with probability $r < q$. A signaling structure is a pair of probabilities that give a chance of signal c or w (correct or wrong) given the true state C or W :

$$\begin{array}{ccccc} & \overset{q}{} & & & \\ & C & W & & \\ c & x & 1-y & \text{and} & \\ w & 1-x & y & & \end{array} \qquad \begin{array}{ccc} \overset{r}{} & & \\ C & W & \\ u & 1-v & . \\ 1-u & v & \end{array} \tag{9}$$

They know there is a proportion $\pi = \frac{27}{74}$ who score 19 or more (with a probability of $q = \frac{193}{200}$ of scoring a question right); and a proportion $\frac{47}{74}$ who scored 19 or less (with a probability of $r = \frac{163}{200}$ of scoring a question right).

Suppose a person got a signal of L; what is the probability that he is a q?

$$
\begin{aligned}
P(q \mid L) &= \frac{P(L \mid q) P(q)}{P(L \mid q) P(q) + P(L \mid r) P(r)} = f(u, v, x, y, q, r, p) \qquad (10)\\
&= \frac{\left(P(L \mid CC) P(CC \mid q) + 2P(L \mid CW) P(CW \mid q) + P(L \mid WW) P(WW \mid q)\right) P(q)}{P(L \mid q) P(q) + P(L \mid r) P(r)} = \\
&= \frac{1}{1 + \frac{\left((1-u)^2 r^2 + 2v(1-u) r(1-r) + v^2 (1-r)^2\right)(1-p)}{\left((1-x)^2 q^2 + 2y(1-x) q(1-q) + y^2 (1-q)^2\right) p}}
\end{aligned}
$$

Also,

$$
\begin{aligned}
P(q \mid M) &= \frac{\left(P(M \mid CC) P(CC \mid q) + 2P(M \mid CW) P(CW \mid q) + P(M \mid WW) P(WW \mid q)\right) P(q)}{P(M \mid q) P(q) + P(M \mid r) P(r)} \qquad (11)\\
&= \frac{1}{1 + \frac{\left(2u(1-u) r^2 + 2(uv + (1-u)(1-v)) r(1-r) + 2v(1-v)(1-r)^2\right)(1-p)}{\left(2x(1-x) q^2 + 2(xy + (1-x)(1-y)) q(1-q) + 2y(1-y)(1-q)^2\right) p}}
\end{aligned}
$$

and finally

$$
\begin{aligned}
P(q \mid H) &= \frac{\left(P(H \mid CC) P(CC \mid q) + 2P(H \mid CW) P(CW \mid q) + P(H \mid WW) P(WW \mid q)\right) P(q)}{P(H \mid q) P(q) + P(H \mid r) P(r)} \qquad (12)\\
&= \frac{1}{1 + \frac{\left(u^2 r^2 + 2u(1-v) r(1-r) + (1-v)^2 (1-r)^2\right)(1-p)}{\left(x^2 q^2 + 2x(1-y) q(1-q) + (1-y)^2 (1-q)^2\right) p}}
\end{aligned}
$$

Of those who got a mark of 0 (that is the equivalent of a low score in BDM, say 3 answers right or less in the sample in BDM) a proportion y^2 got signal L, a proportion $2y(1-y)$ got signal M and a proportion $(1-y)^2$ got signal H. Similarly we can calculate the distribution over signals for people with marks of 1 and 2, and for types r (ommitted):

Probability of each signal given score (for a type q).

mark ↓ /signal →	L	M	H
0	y^2	$2y(1-y)$	$(1-y)^2$
1	$y(1-x)$	$xy + (1-x)(1-y)$	$x(1-y)$
2	$(1-x)^2$	$2x(1-x)$	x^2

(13)

Then, the probability of scoring 19 or *A*bove given each signal $s = L, M, H$ is

$$
\begin{aligned}
P(A \mid s) &= P(A \mid q) P(q \mid s) + P(A \mid r) P(r \mid s)\\
&= \left(20q^{19}(1-q) + q^{20}\right) P(q \mid s) + \left(20r^{19}(1-r) + r^{20}\right)(1 - P(q \mid s))
\end{aligned}
$$

Now the question is how to map this into the data and tests. The nice thing about the model is that there is no need to assume "noise" (like people wanted to say 55 but said 60),

but in order to avoid introducing noise, we need as many signals as numbers of confidence were mentioned (or, say, 13 categories of 5% which were mentioned: confidence 0-4%; 5-9%; etc. (not all ranges were mentioned). In what follows we will assume that there were only three declarations of confidence: 0-50% (inclusive); 51-79%; 80-100%. This is to develop the tests.

The data from BDM experiment 2 then becomes

	Score < 19			Score ≥19			Total		
mark ↓ /bet →	$0-50$	$51-79$	$80-100$	$0-50$	$51-79$	$80-100$	$0-50$	$51-79$	$80-100$
0	9	3	4	1	3	1	10	6	5
1	3	8	7	1	1	2	4	9	9
2	5	4	4	1	9	8	6	13	12

(14)

2.1.1 One signalling structure

We now calculate the likelihood of this sample using only one signalling structure for both types of people (the skilled and unskilled); that is, we set $u = x$ and $v = y$ (in the signalling structures in 9). In order to do that, we assume we know what signal the individual observed (a H signal corresponds to a high bet; a M signal to an intermediate bet; and a L signal to a low bet), and we ask ourselves what is the chance that the person observed that signal s_i, given his mark m_i in the sample questions. Then the likelihood of the sample (given 13 and 14) is: (for number of people n_{ij} who got mark i and signal j)

$$\begin{aligned}
\prod_{i=1}^{74} P(s_i \mid m_i) &= \left(y^2\right)^{n_{0L}} \left(2y(1-y)\right)^{n_{0M}} \left((1-y)^2\right)^{n_{0H}} \left(y(1-x)\right)^{n_{1L}} \left(xy+(1-x)(1-y)\right)^{n_{1M}} \left(x(1-y\right.\\
&= 2^{n_{0M}+n_{2M}} y^{2n_{0L}+n_{0M}+n_{1L}} (1-y)^{2n_{0H}+n_{0M}+n_{1H}} x^{n_{1H}+n_{2M}+2n_{2H}} (1-x)^{2n_{2L}+n_{1L}+n_{2M}} (1-x\\
&= \left(y^2\right)^{10} \left(2y(1-y)\right)^{6} \left((1-y)^2\right)^{5} \left(y(1-x)\right)^{4} \left(xy+(1-x)(1-y)\right)^{9} \left(x(1-y)\right)^{9} \left((1-x\right.\\
&= y^{20} \left(2y(1-y)\right)^{6} (1-y)^{10} y^4 (1-x)^4 \left(xy+(1-x)(1-y)\right)^{9} x^9 (1-y)^9 (1-x)^{12} (2x(\\
&= 524\,288 x^{46} y^{30} (x-1)^{29} (y-1)^{25} (1-x-y+2xy)^9
\end{aligned}$$

In order to find the x and y that maximize the likelihood we do two things.

1. Pick a new prior, a new chance of being of type q. In order to use a better prior (we had rejected that the data came from a rational model in BDM, so we will pick an incorrect prior, and then assume that the signalling structure, and the updating, is "rational"), consider the following changes in the above model.

- The actual average score of those in the top half (counting 10 people who scored 18) was $\frac{701}{37}$; for those at the bottom, it was $\frac{584}{37}$. People reported an average probability of being top half of 67%. If this had been the actual proportion of people who would score on average $\frac{701}{37}$, then the actual average score would have been
$$\frac{67}{100}\frac{701}{37}+\frac{33}{100}\frac{584}{37}=\frac{66\,239}{3700}=17.902$$
instead of $\frac{1285}{74}=\frac{1}{2}\frac{701}{37}+\frac{1}{2}\frac{584}{37}=17.365$.

  Then, we find the prior in the current model so that the average score in the model is this "expected" score (with the distorted prior):
$$20\left(nq+(1-n)r\right)=\frac{66\,239}{3700}\Leftrightarrow 20\left(n\frac{193}{200}+(1-n)\frac{163}{200}\right)=\frac{66\,239}{3700}\Leftrightarrow n=\frac{5929}{11\,100}=0.53414.$$
  That is: we will check the test with this prior (the correct one is $\frac{27}{74}$).

- Since the median score was 18, we assign 10 of those who scored 18 to the top half, and the rest to the bottom half, so that the top half is 10 who scored 18, 19 who scored 19 and 8 who scored 20. Then, since the average probability reported in the experiment of being in the top half was 67%, the reported probability (assuming they knew that 10 out of 37 in the top half would not score 19 or 20), the reported probability of scoring 19 or 20 was $67*\frac{27}{37}=\frac{1809}{37}=48.892$.

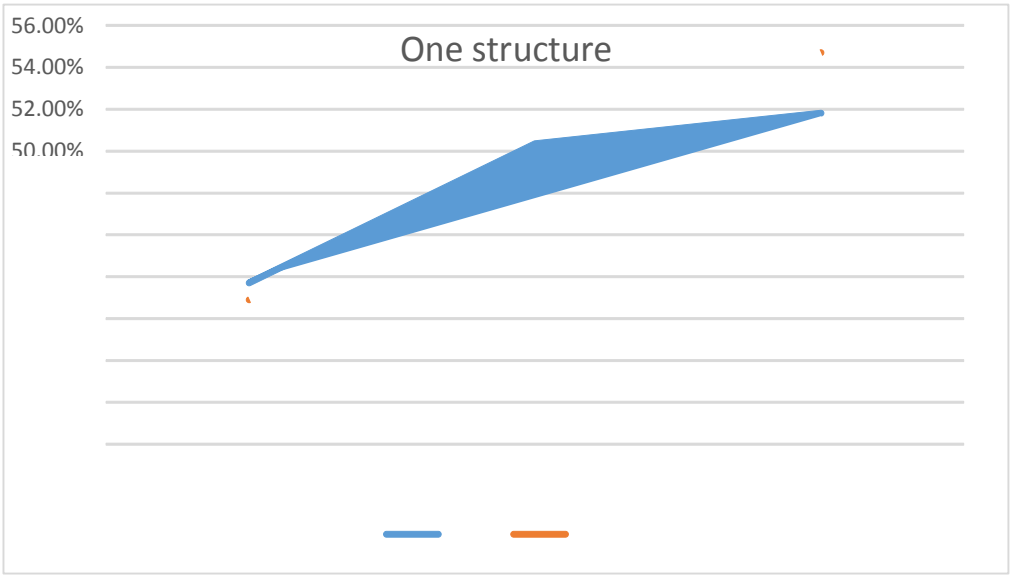
2. Maximize the likelihood of observing the sample we observed (the likelihood in equation 15), conditional on $P(q\mid L)\leq 50*\frac{27}{37}=36.486$, $P(q\mid M)\in[50,80]*\frac{27}{37}=[36.486,58.378]$ and $P(q\mid H)\geq 80*\frac{27}{37}=58.378$ (given in equations 10, 11 and 12), since we assume that when people observe those signals, they report posteriors in those ranges.

These calculations yield : $(x,y)=(0.8345,0.6898)$. Then, $P(q\mid L)=36\%$, $P(q\mid M)=0.470\,53$ and $P(q\mid H)=0.58405$. Also, the reported *A*verage probabilities of being a type q, conditional on a mark of $0,1$ or 2 are (still with $u=x,v=y$) :

$$\begin{aligned}
A(0) &= P(q\mid L)P(L\mid 0)+P(q\mid M)P(M\mid 0)+P(q\mid H)P(H\mid 0) \qquad (16)\\
&= P(q\mid L)y^2+P(q\mid M)2y(1-y)+P(q\mid H)(1-y)^2\\
A(1) &= P(q\mid L)P(L\mid 1)+P(q\mid M)P(M\mid 1)+P(q\mid H)P(H\mid 1)\\
&= P(q\mid L)y(1-x)+P(q\mid M)(xy+(1-x)(1-y))+P(q\mid H)x(1-y)\\
A(2) &= P(q\mid L)(1-x)^2+P(q\mid M)2x(1-x)+P(q\mid H)x^2
\end{aligned}$$

With the calculated x and y, we obtain $A(0) = 0.428\,85$, $A(1) = 0.487\,30$ and $A(2) = 0.54656$. For the data in BDM, people who answered 5 correct said their average chance of being in the top half was 51.55% (71% times 27/37). For those who anwered 4, it was 50.6 (69(27/37), and for those who anwered 3, it was 43.7 (60 times 27/37) or $49.6 = 68 * \frac{27}{37}$ if we consider those who answered 3 or less (those who answered 0 were probably not dumb, but actually probably better than those who answered 3 sample questions correctly).

This yields the following comparison between the actual data, and that predicted by the model



The model seems to perform pretty well, but the whole point of this section to actually run a statistical test. For that we need to calculate the best possible model where skilled and unskilled have different signalling structures. We now turn to that, but the bottom line will be that we do not reject one structure; the p value is 18%.

2.1.2 Two signalling structures.

Now assume there are two signalling structures, one for the skilled and one for the unskilled. For the Unskilled we have let u and v and for the Skilled, we keep x and y. In this case, we let b_{ij} denote the number of skilled people with mark i and signal j; similarly, a_{ij} for unskilled people. The likelihood is then

$$\begin{aligned}
\prod_{i=1}^{27} P_S(s_i \mid m_i) \prod_{i=28}^{74} P_U(s_i \mid m_i) &= \left(y^2\right)^{b_{0L}} \left(2y(1-y)\right)^{b_{0M}} \left((1-y)^2\right)^{b_{0H}} \left(y(1-w)\right)^{b_{1L}} (xy + (1-x)(1 \\
&\quad \left(v^2\right)^{a_{0L}} \left(2v(1-v)\right)^{a_{0M}} \left((1-v)^2\right)^{a_{0H}} \left(v(1-u)\right)^{a_{1L}} (uv + (1-u)(1 \\
&= \left(y^2\right) \left(2y(1-y)\right)^3 \left((1-y)^2\right) \left(y(1-x)\right) (xy + (1-x)(1-y)) (x(1 \\
&\quad \left(v^2\right)^9 \left(2v(1-v)\right)^3 \left((1-v)^2\right)^4 \left(v(1-u)\right)^3 (uv + (1-u)(1-v))^8 (\\
&= 4096 x^{27} y^6 (x-1)^{12} (y-1)^7 (x+y-2xy-1)\, 128 u^{19} v^{24} (1-u)^{17} (
\end{aligned}$$

We impose the same restrictions as before in terms of the posteriors being of the size they should, $P(q \mid L) \leq 36.486$, $P(q \mid M) \in [36.486, 58.378]$ and $P(q \mid H) \geq 58.378$ (given in equations 10, 11 and 12), and maximize this likelihood to obtain $(u, v, x, y) = (0.8318, 0.7009, 0.8417, 0.6598)$

We then obtain, $P(q \mid L) = 0.332\,58$, $P(q \mid M) = 0.459\,33$ and $P(q \mid H) = 0.591\,58$. Also, we need to re calculate the As (the average reported posterior of people who score 0, 1 or 2):

$$\begin{aligned}
A(0) &= A(0;q) P(q) + A(0;r) P(r) \\
&= \left[P(q \mid L) P(L \mid 0,q) + P(q \mid M) P(M \mid 0,q) + P(q \mid H) P(H \mid 0,q)\right] P(q) + \\
&\quad \left[P(q \mid L) P(L \mid 0,r) + P(q \mid M) P(M \mid 0,r) + P(q \mid H) P(H \mid 0,r)\right] P(r) \\
&= \left[P(q \mid L) y^2 + P(q \mid M) 2y(1-y) + P(q \mid H)(1-y)^2\right] p + \\
&\quad \left[P(q \mid L) v^2 + P(q \mid M) 2v(1-v) + P(q \mid H)(1-v)^2\right] (1-p) \\
A(1) &= \left[P(q \mid L) y(1-x) + P(q \mid M)(xy + (1-x)(1-y)) + P(q \mid H) x(1-y)\right] p + \\
&\quad \left[P(q \mid L) v(1-u) + P(q \mid M)(uv + (1-u)(1-v)) + P(q \mid H) u(1-v)\right] (1-p) \\
A(2) &= \left[P(q \mid L)(1-x)^2 + P(q \mid M) 2x(1-x) + P(q \mid H) x^2\right] p + \\
&\quad \left[P(q \mid L)(1-u)^2 + P(q \mid M) 2u(1-u) + P(q \mid H) u^2\right] (1-p)
\end{aligned}$$

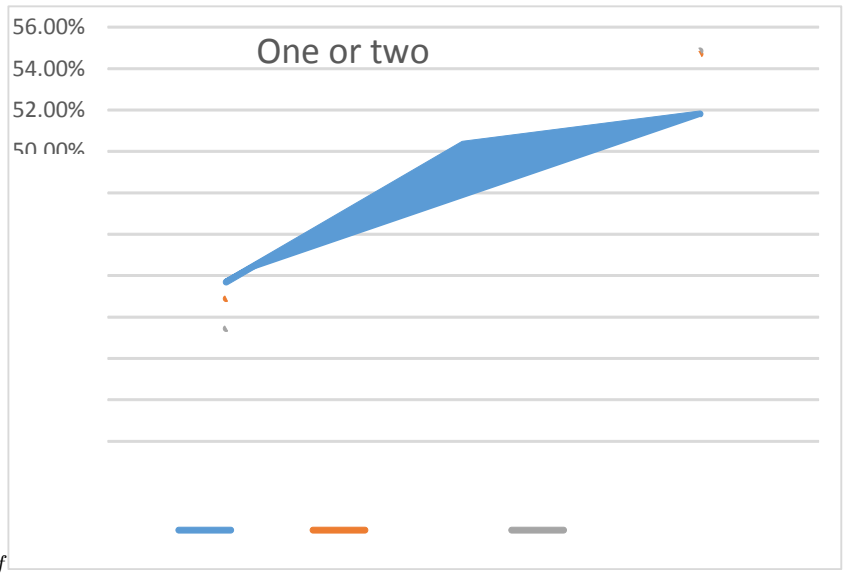
The results are $A(0) = 0.414\,54$, $A(1) = 0.480\,85$ and $A(2) = 0.548\,64$.

For the skilled, the average reported probability of being in the top half, As, if they obtain 0,1 or 2 right are exactly as in (16), while for unskilled the expressions $Au(\cdot)$ are the same expressions, but with u instead of x and v instead of y.

	0	1	2
As	0.41946	0.483 96	0.549 85
Au	0.408 89	0.477 29	0.547 25

The following graph illustrates the predictions of one model (previous section), of two models (this section) and the actual data.

and two averages

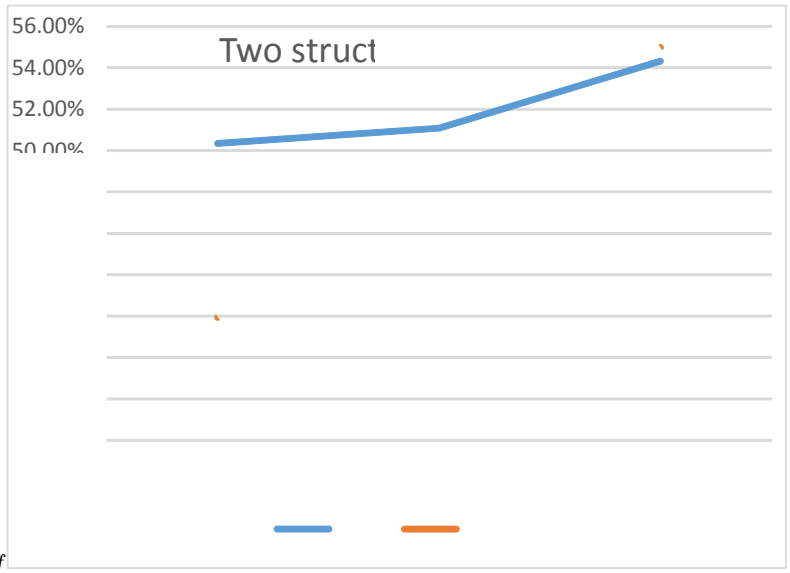


3.pdf

The two models seems to do worse, but that is just for averages; we are optimizing something different. Still, two models does not seem to be a large improvement over one.

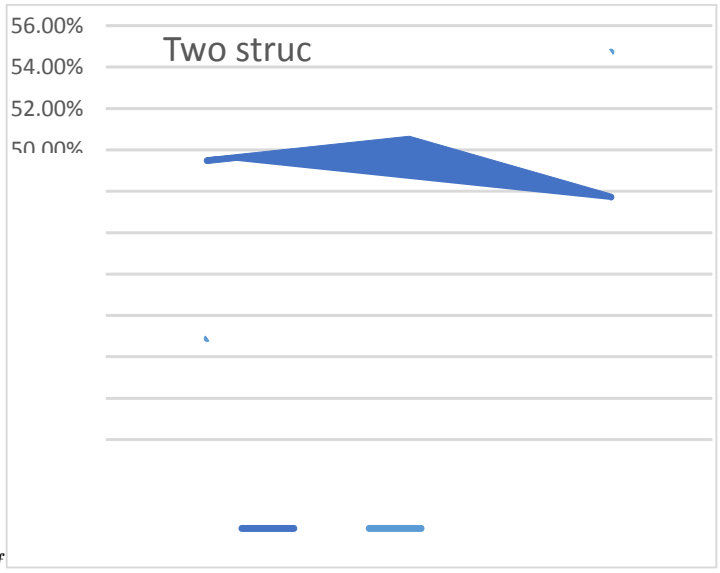
As for the structures and data for each group, we have

Skilled



4.pdf

Unskilled



5.pdf

2.1.3 Actual Test

Let us call a sample the first two panels of (14); the third is just the sum. Let us say that H_0 is that both signalling structures are the same, and $(x, y) = (0.8345, 0.6898)$. Also, H_1 is one signalling structure for each group, and $(v, u, w, z) = (0.8318, 0.7009, 0.8417, 0.6598)$. Let

$$L(n_{ij}; x, y) = 2^{n_{0M}+n_{2M}} y^{2n_{0L}+n_{0M}+n_{1L}} (1-y)^{2n_{0H}+n_{0M}+n_{1H}} x^{n_{1H}+n_{2M}+2n_{2H}} (1-x)^{2n_{2L}+n_{1L}+n_{2M}} (1-x-y+2$$

be the likelihood of the sample, so that for a sample X we have

$$\Lambda(X) \equiv \frac{L(X \mid H_0)}{L(X \mid H_1)} \equiv \frac{L(n_{ij}; x, y)}{L(a_{ij}; u, v) L(b_{ij}; w, z)} \tag{17}$$

and using $n_{ij} = a_{ij} + b_{ij}$ we obtain (let us abuse notation and $a_{ij} = a_j^i$ (for horizontal space)

$$\Lambda = \left|\frac{y}{v}\right|^{2a_L^0+a_M^0+a_L^1} \left|\frac{y}{z}\right|^{2b_L^0+b_M^0+b_L^1} \left|\frac{1-y}{1-v}\right|^{2a_H^0+a_M^0+a_H^1} \left|\frac{1-y}{1-z}\right|^{2b_H^0+b_M^0+b_H^1} \left|\frac{x}{u}\right|^{a_H^1+a_M^2+2a_H^2} \left|\frac{x}{w}\right|^{b_H^1+b_M^2+2b_H^2} \left|\frac{1-x}{1-u}\right|^{2a_L^2-}$$

The idea of the test we will run is the following. The likelihood ratio statistic $\Lambda(X)$ is a function of the sample X; the likelihood ratio for the sample we observed is less than 1, since in the denominator we optimize over two sets of parameters which can be set equal to x and y (in the numerator), so the likelihood in the denominator is larger than that on the numerator. The question is whether the one we obtained for our particular sample is "too low"; if it is, then one model is not good enough to represent the data we have. What does too low mean? It means that (under the null that the signalling structures are equal) it is unlikely that most samples have a larger likelihood ratio.

The test we describe above is the Neyman-Pearson test, which requires that we know the distribution of Λ; but the only random thing in Λ is X. So we need to know what is the probability that out of $\sum_{ij} a_{ij} + \sum_{ij} b_{ij}$ individuals, we will have a_{ij} being unskilled and obtaining mark i and signal j, and b_{ij} being skilled, obtaining mark i and observing signal j. That is very simple, it is a multinomial (using H_0). We need to calculate the probability that 74 subjects fall into one of 18 categories: $\{\theta = S, U\} \times \{m = 0, 1, 2\} \times \{s = L, M, H\}$. The chance of an individual falling on each category are given by

Unskilled U,.with prob $\frac{47}{74}$

mark $\downarrow$ /signal $\rightarrow$	L	M	H
0 w $P(0 \mid U) = \left(1 - \frac{163}{200}\right)^2$	$\frac{47}{74}\left(1 - \frac{163}{200}\right)^2 y^2$	$\frac{47}{74}\left(1 - \frac{163}{200}\right)^2 2y(1-y)$	$\frac{47}{74}\left(1 - \frac{163}{200}\right)^2 (1-y)^2$
1 w $P(1 \mid U) = 2 * \frac{163}{200}\left(1 - \frac{163}{200}\right)$	$\frac{47}{74}\frac{6031}{20\,000} y(1-x)$	$\frac{47}{74}\frac{6031}{20\,000}(2xy - y - x + 1)$	$\frac{47}{74}\frac{6031}{20\,000} x(1-y)$
2 w $P(2 \mid U) = \left(\frac{163}{200}\right)^2$	$\frac{47}{74}\left(\frac{163}{200}\right)^2 (1-x)^2$	$\frac{47}{74}\left(\frac{163}{200}\right)^2 2x(1-x)$	$\frac{47}{74}\left(\frac{163}{200}\right)^2 x^2$

(18)

Skilled S,.with prob $\frac{27}{74}$

mark $\downarrow$ /signal $\rightarrow$	L	M	H
0 w $P(0 \mid S) = \left(1 - \frac{193}{200}\right)^2$	$\frac{27}{74}\left(1 - \frac{193}{200}\right)^2 y^2$	$\frac{27}{74}\left(1 - \frac{193}{200}\right)^2 2y(1-y)$	$\frac{27}{74}\left(1 - \frac{193}{200}\right)^2 (1-y)^2$
1 w $P(1 \mid S) = 2 * \frac{193}{200}\left(1 - \frac{193}{200}\right)$	$\frac{27}{74}\frac{1351}{20\,000} y(1-x)$	$\frac{27}{74}\frac{1351}{20\,000}(2xy - y - x + 1)$	$\frac{27}{74}\frac{1351}{20\,000} x(1-y)$
2 w $P(2 \mid S) = \left(\frac{193}{200}\right)^2$	$\frac{27}{74}\left(\frac{193}{200}\right)^2 (1-x)^2$	$\frac{27}{74}\left(\frac{193}{200}\right)^2 2x(1-x)$	$\frac{27}{74}\left(\frac{193}{200}\right)^2 x^2$

Denote the probability that an individual is of type $\theta \in \{U, S\}$ obtains mark m and observes signal s by θ_{ms} (these numbers come from the matrices (18) above). Then the probability of a sample $X = (x_1, ..., x_{18}) = (a_{0L}, a_{0M}, a_{0H}, a_{1L}, ..., a_{2H}, b_{0L}, b_{0M}, ..., b_{2H})$ is a standard multinomial distribution:

$$P(X) = \frac{74!}{\prod_1^{18} x_i!} U_{0L}^{x_1} U_{0M}^{x_2} U_{0H}^{x_3} U_{1L}^{x_4} ... U_{2H}^{x_9} S_{0L}^{x_{10}} S_{0M}^{x_{11}} ... S_{2H}^{x_{18}}.$$

With this distribution, we can calculate the probability of each X (arising from a sample of 74 people); for each X, we can calculate the value of $\Lambda(X)$. Then, we can calculate the probability that $\Lambda \leq \eta$ for each η. So we fix η^* so that $P(\Lambda \leq \eta^*) = 5\%$, and check whether the sample we acually observed had $\Lambda(X) \leq \eta^*$.

The result of running this test is that the probability that $P(\Lambda \leq \Lambda(X))$ (the probability of Λ falling below the one we found for our sample) is 18.3%. That is, we cannot reject one signalling model for all.

As for work in progress, we will analyze with the data of this experiment, whether the signalling structure of the skilled seems to be less noisy than that of the unskilled. That can be analyzed with the estimates of v, u, w, z.

2.1.4 Related literature

Several authors have proposed the idea the mean reversion can explain the stylized fact we study in this section. Ehrlinger et al. adress their critiques, but they don't perform a statistical test. They first estimate a β in a regression, and then they correct it with a parameter reflecting the reliability of the measurement of the variables involved, but the analysis has no theoretical basis, and in the end the analysis consists of a visual comparison of two regressions (see study 1, the summary, and figure 2):

quote

> Figure 2 depicts the results of an analysis in which perceived performance (including, separately, perceived mastery of course material, percentile score and raw score) is regressed on objective performance. It also depicts what the regression analysis looks like after assuming perfect reliability. As seen in the figure, across three different measures of perceived performance, the relationship between perceived and actual performance was stronger once unreliability was corrected for, but this strengthening was minimal. For example, in terms of test performance relative to other students, participants in the bottom quartile typically overestimated their percentile rank by 49 percentile points. After correcting for unreliability, their overestimates are reduced by only roughly 5 points. In terms of raw score, bottom performers overestimated their score by 8.4 points before correction; 7.2 points after. However, as Figure 2 also shows, a good portion of the misestimates among top performers were eliminated when we corrected for unreliability. For example, concerning estimates of raw score, top performers underestimated their score by 1.7 points before correction; but only .2 points afterward.

6.pdf

2.2 Experiment 4, Unskilled and Unaware, new data.

Experiment 4 was run together with Experiments 2 and 5, in Amsterdam, with 313 subjects. It is very similar to the analysis in the previous section concerning Experiment 3, but the results are different: we reject one signaling structure.

Subjects answer three sample questions, they are told that the test they are about to take is similar in difficulty to the sample questions, they are told that in populations similar to theirs, the better performing half usually answers 7 or more correctly out of the 12 questions, and then they have to state what they think is their chance of being in the top half. Then they take the test, and they are ranked in two halves.

We now describe a general way of doing a likelihood test given our data, it is more abstract than the previous section, where signals received depended on whether individuals had answered each question in the sample correctly, or not. Still, to be clear, the difference

in the results is not due to the different model (we tested the new data with the model of the previous section, and the same results emerged, as with this model). We define θ_S as the skilled, those who scored in the top half of test takers, and θ_U as those who did not.

If we were going to use all the available data, we would need to estimate $4 \times 100 \times 2 = 800$ parameters, corresponding to the chances of declaring $0 - 100\%$ chance of being in the top half for 2 types of subjects, who score $0 - 3$ in the sample questions. Since it would be too demanding in terms of data collection to have enough subjects to run tests with so many parameters, we group subjects' declarations into four intervals: those who declare their chances of being in the top half to be less than 25%, between 25 and 50%, between 50 and 75%, and above 75%. Our model then estimates $4 \times 4 \times 2$ parameters. We let $N = 4$, and $z_i = i/N$ for $i = 0, ..., N$. This defines N intervals, $[0, z_1], [z_1, z_2], ..., [z_{N-1}, 1]$. One interpretation of KD is that they say that, conditonal on whatever score they got on the sample questions (they don't know their score, but receive signals from their performance), unskilled people are unable to figure out if they are good or bad, while the skilled can infer their ability better. This would mean that for each score $m_i \in \{0, 1, 2, 3\}$ there is one information structure for the skilled and another for the unskilled, where the information of the skilled is better than that of the unskilled. More precisely, for each m_i there is a distribution $p^i \in \Delta^{N-1}$ for the skilled (say, p^i_j is the probability of observing s_j, where $P(\theta_S \mid s_j) \in [z_{j-1}, z_j]$ is the estimated probability of placing in the top half of a person who scored m_i, and who is actually skilled, in the sense of having scored in the top half in the test), and a distribution $q^i \in \Delta^{N-1}$ for the unskilled (q^i_j is the probability of observing s_j for an unskilled individual who scored m_i). Let μ^S be the distribution over marks of the skilled (say, μ^S_1 is the chance with which the skilled answer 1 sample question correctly), and similarly μ^U is the distribution of the unskilled.

In our data, 111 subjects scored below the median, 57 exactly the median of 7/12, and 145 above the median. We assign 45 of those who scored at the median to the unskilled and the rest to the skilled, so that there are 156 unskilled and 157 skilled.[5] With this procedure,

[5] In order to avoid introducing noise in the estimation (depending on which individuals were assigned to each group), we proceeded as follows. For each of the three groups, we computed the average distribution over $\{s_i\}_1^N$, conditional on each mark in the sample. The distribution over signals of the unskilled with a given mark was the weighted average of: the distribution of those who scored below the median and had that mark; and that of those who scored the median and had that mark. The weights were: the number of subject who scored below the median; and the number of subjects who scored at the median multiplied by

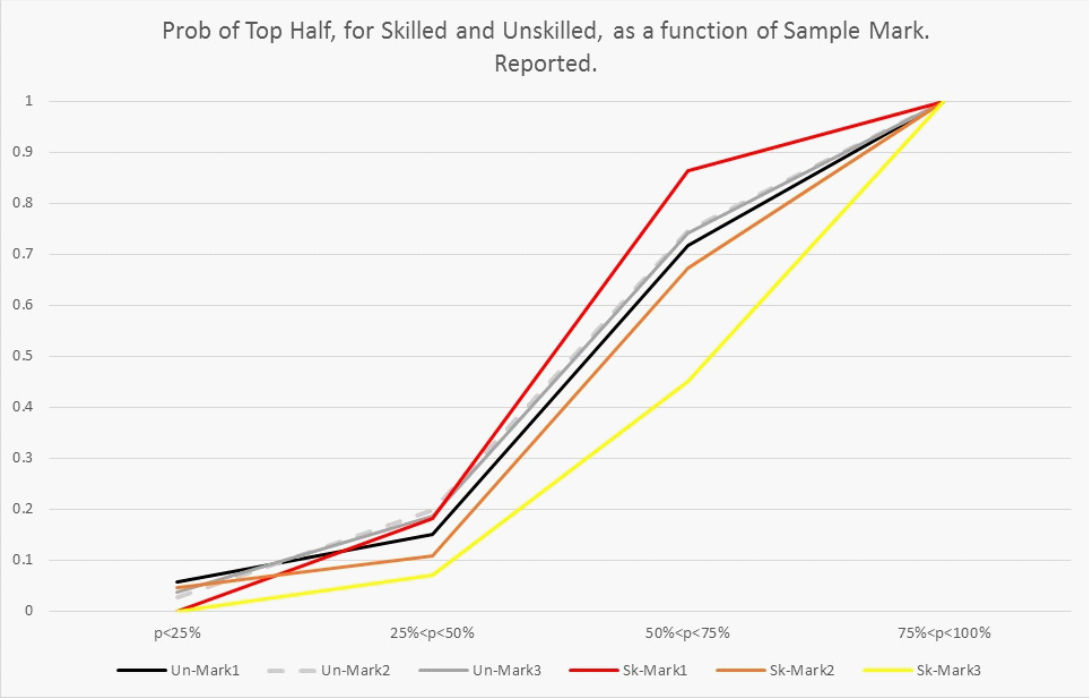
the sample we observed in terms of marks and self placements was:

	Types θ_U					Types θ_S				
mark ↓ /bet →	s_1	s_1	s_3	s_4	μ^U	s_1	s_2	s_3	s_4	μ^S
0	20	20	60	0	3.2%	0	0	0	0	0%
1	6	9	57	28	33.6%	0	19	68	14	14.3%
2	3	17	55	25	45.6%	4	6	56	33	40.7%
3	3	13	57	28	17.5%	0	8	38	54	45.0%
					156					157

(19)

Distribution over signals for each type, conditional on mark in sample questions.

Before getting into the model, it is worth noting that indeed the unskilled seem unable to judge how well they will do: ommitting those who scored 0, which are very few, the distribution over declarations of chances of being in the top half (over the four categories, < 25, between 25 and 50, between 50 and 75, and over 75) seem to be independent of their mark in the sample questions, while that is not true for the skilled. For the skilled, the predictions are almost ordered by first order stochastic dominance, with those who score 1 making the lowest predictions, and those who scored 3 the highest.

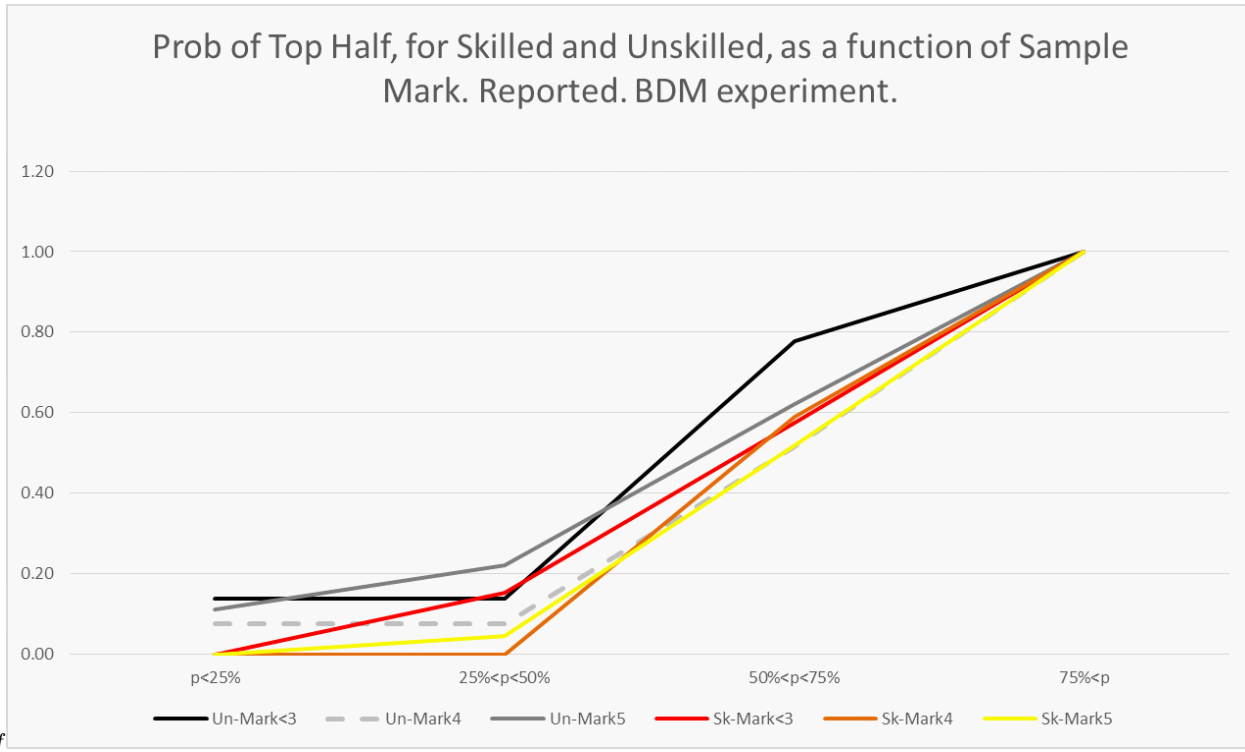


In contrast, in the BDM experiment, Experiment 3, the unskilled seem to be good at estimating their chances of being in the top half: notice that the unskilled who score less

45/57. Similarly for the skilled.

than 3 in the sample have the lowest predictions of being in the top half, followed by the unskilled who score 5 in the sample.

experiment



8.*pdf*

Let a_j^i denote the number of unskilled subjects who score i, and observe signal s_j, and let b_j^i be the equivalent number for the skilled. The likelihood of a sample is then

$$L\left(a_j^i, b_j^i, p, q\right) = \prod_{j=1}^{N}\left(q_j^0\right)^{a_j^0} ... \prod_{j=1}^{N}\left(q_j^3\right)^{a_j^3} \prod_{j=1}^{N}\left(p_j^0\right)^{b_j^0} ... \prod_{j=1}^{N}\left(p_j^3\right)^{b_j^3}. \tag{20}$$

The basic question we want to ask is how much more likely is the sample if we allow for different signalling structures for skilled and unskilled, versus, if we restrict them to have the same information structure?; how much better does this two-structure model explain the data?

In order to answer it we need to compare the likelihood with the best (in terms of maximizing the likelihood) p's and q's in the two-structure model with the likelihood with the best signaling structure (r) in the one-structure model. We need to choose $p^i, q^i \in \Delta^N$, $i = 0, ..., 3$, (so that each $\sum_j p_j^i = \sum_j q_j^i = 1$) to maximize (20) subject to the constraint that for each j an individual who observes signal j wants to declare a probability of being

in the top half of $P(\theta_S \mid s_j) \in [z_{j-1}, z_j]$

$$z_{j-1} \leq P(\theta_S \mid s_j) = \frac{\pi \sum_{i=0}^{3} \mu_i^S p_j^i}{\pi \sum_{i=0}^{3} \mu_i^S p_j^i + (1-\pi) \sum_{i=0}^{3} \mu_i^U q_j^i} \leq z_j \Leftrightarrow \frac{1-z_{j-1}}{z_{j-1}} \geq \frac{1-\pi}{\pi} \frac{\sum_{i=0}^{3} \mu_i^U q_j^i}{\sum_{i=0}^{3} \mu_i^S p_j^i} \geq \frac{1-z_j}{z_j}.$$

We also choose the prior, since we know (see BDM, and that is also the case in this experiment) that $\pi = \frac{1}{2}$ is rejected by the data. We choose a distorted prior, and then let people be rational in their updating. We set the prior equal to the average reported probability of being in the top half. Call $\left\{\overline{p}^i\right\}$ and $\left\{\overline{q}^i\right\}$ the solution to the problem.

This case of two signaling structures is an interpretation of KD, and in this context, the claim of KD would be that a model where skilled and unskilled process information in the same way would be rejected by the data. That is, if we set $p^i = q^i$ for all i in (20) and maximize the likelihood. Concretely, we need to choose $p^i \in \Delta^{N-1}$, $i = 0, ..., 3$, (so that each $\sum_j p_j^i = 1$) to maximize

$$\begin{aligned} L\left(a_j^i, b_j^i, p, p\right) &= \prod_{j=1}^{N} \left(p_j^0\right)^{a_j^0+b_j^0} ... \prod_{j=1}^{N} \left(p_j^3\right)^{a_j^3+b_j^3} \text{ subject to, for all } j \qquad (21) \\ z_{j-1} &\leq P(\theta_S \mid s_j) = \frac{\pi \sum_{i=0}^{3} \mu_i^S p_j^i}{\pi \sum_{i=0}^{3} \mu_i^S p_j^i + (1-\pi) \sum_{i=0}^{3} \mu_i^U p_j^i} \leq z_j \Leftrightarrow \\ \frac{1-z_{j-1}}{z_{j-1}} &\geq \frac{1-\pi}{\pi} \frac{\sum_{i=0}^{3} \mu_i^U p_j^i}{\sum_{i=0}^{3} \mu_i^S p_j^i} \geq \frac{1-z_j}{z_j}. \end{aligned}$$

Let $\{\widetilde{p}^i\}$ be the solution to this problem.

To be clear, in this interpretation the skilled and unskilled receive different information (even in the case in which they have the same information structures) because they score differently in the sample questions. This seems to be a fair description of the theory KD try to disprove: skilled and unskilled make different predictions because they find the sample questions differentially hard; but conditional on how they do, they infer the same thing from the same signal.

The question is now whether the solution to (20) improves "a lot" relative to the solution to (21). As with Experiment 3, the idea of the test is to calculate a statistic $\Lambda(X)$, a function of the sample X that will be small if one signalling structure is not good; then assuming the true model is with one signalling structure, we will calculate the probability that "a lot" of samples $\widetilde{X}$ yield statistics which are larger than $\Lambda(X)$, in which case we will conclude that one model is not a good enough representation of the data.

Concretely, recall that $\left\{\overline{p}^i\right\}$ and $\left\{\overline{q}^i\right\}$ was the solution to the problem in (20) with two signaling structures, and that $\{\widetilde{p}^i\}$ was the solution to the problem in (21) with just one

signaling structure. Recall also that a sample X is described by $\left\{a_j^i\right\}, \left\{b_j^i\right\}$ where a_j^i denotes the number of unskilled subjects who score i, and observe signal j, and b_j^i is the equivalent number for the skilled. Then, define for a sample X, and H_0 ="there is just one signaling structure $\{\widetilde{p}^i\}$",

$$\Lambda(X) = \frac{L(X \mid H_0)}{L(X \mid H_1)} = \frac{L\left(a_j^i, b_j^i; \widetilde{p}, \widetilde{p}\right)}{L\left(a_j^i, b_j^i; \overline{p}, \overline{q}\right)} = \frac{\prod_{j=1}^N \left(\widetilde{p}_j^0\right)^{a_j^0+b_j^0} ... \prod_{j=1}^N \left(\widetilde{p}_j^3\right)^{a_j^3+b_j^3}}{\prod_{j=1}^N \left(\overline{q}_j^0\right)^{a_j^0} ... \prod_{j=1}^N \left(\overline{q}_j^3\right)^{a_j^3} \prod_{j=1}^N \left(\overline{p}_j^0\right)^{b_j^0} ... \prod_{j=1}^N \left(\overline{p}_j^3\right)^{b_j^3}}.$$

The (exact) Neyman-Pearson test requires that we know the distribution of Λ; but the only random thing in Λ is X. So we need to know what is the probability that out of $\sum_{ij} a_j^i + \sum_{ij} b_j^i$ individuals, we will have a_j^i being unskilled and obtaining mark i and signal j, and b_j^i being skilled, obtaining mark i and observing signal j. That is very simple, it is a multinomial (using H_0). We need to calculate the probability that 313 subjects fall into one of $8N$ categories: $\{\theta = S, U\} \times \{m = 0, 1, 2, 3\} \times \{s = s_1, ..., s_N\}$. Then, recalling that μ^S is the distribution over marks of the skilled, and μ^U of the unskilled, the chances of an individual falling on each category are given by

$$\begin{array}{cccccc|cccccc}
 & \multicolumn{5}{c|}{\text{Unskilled } U, \text{.with prob } 1-\pi} & & \multicolumn{5}{c}{\text{Skilled } S, \text{.with prob } \pi} \\
 & s_0 & & \cdots & & s_N & & s_0 & & \cdots & & s_N \\
m_0 & (1-\pi)\,\mu_0^U \widetilde{p}_0^0 & & \cdots & & (1-\pi)\,\mu_0^U \widetilde{p}_N^0 & m_0 & \pi \mu_0^S \widetilde{p}_0^0 & & \cdots & & \pi \mu_0^S \widetilde{p}_N^0 \\
m_1 & (1-\pi)\,\mu_1^U \widetilde{p}_0^1 & & \cdots & & (1-\pi)\,\mu_1^U \widetilde{p}_N^1 & m_1 & \pi \mu_1^S \widetilde{p}_0^1 & & \cdots & & \pi \mu_1^S \widetilde{p}_N^1 \\
m_2 & (1-\pi)\,\mu_2^U \widetilde{p}_0^2 & & \cdots & & (1-\pi)\,\mu_2^U \widetilde{p}_N^2 & m_2 & \pi \mu_2^S \widetilde{p}_0^2 & & \cdots & & \pi \mu_2^S \widetilde{p}_N^2 \\
m_3 & (1-\pi)\,\mu_3^U \widetilde{p}_0^3 & & \cdots & & (1-\pi)\,\mu_3^U \widetilde{p}_N^3 & m_3 & \pi \mu_3^S \widetilde{p}_0^3 & & \cdots & & \pi \mu_3^S \widetilde{p}_N^3
\end{array} \tag{22}$$

Denote the probability that an individual is of type $\theta \in \{U, S\}$ obtains mark m and observes signal s by θ_{ms} (these numbers come from the matrices (22) above). Then the probability of a sample $X = (x_1, ..., x_{8N}) = (a_1^0, a_2^0, ..., a_N^0, a_1^1, ..., a_N^1, ..., a_N^3, b_1^0, b_2^0, ..., b_N^3)$ is a standard multinomial distribution:

$$\begin{aligned}
P(X) &= \frac{313!}{\prod_1^{8N+8} x_i!} U_{01}^{x_1} U_{02}^{x_2} ... U_{0N}^{x_N} U_{11}^{x_{N+1}} U_{12}^{x_{N+2}} ... U_{3N}^{x_{4N}} S_{01}^{x_{4N+1}} S_{02}^{x_{4N+2}} ... S_{3N}^{x_{8N}} \\
&= \frac{313!}{\prod a_j^i! \prod b_j^i!} U_{01}^{a_1^0} U_{02}^{a_2^0} ... U_{0N}^{a_N^0} U_{11}^{a_1^1} U_{12}^{a_2^1} ... U_{3N}^{a_N^3} S_{01}^{b_1^0} S_{02}^{b_2^0} ... S_{3N}^{b_N^3}.
\end{aligned}$$

With this distribution, we can calculate the probability of each X (arising from a sample of 313 people); for each X, we can calculate the value of $\Lambda(X)$. Then, we can calculate the

probability that $\Lambda \leq \eta$ for each η. So we fix η^* so that $P(\Lambda \leq \eta^*) = 5\%$, and check whether the sample we acually observed had $\Lambda(X) \leq \eta^*$.

We simulated this distribution, with ten million draws, several times, and in every case proportion of Λs lower than the one in the experiment was less than one tenth of 1%. The data reject one signalling structure at all conventional confidence levels.

To check whether the difference between Experiment 3 (do not reject one signaling structure with a p value of 18%) and Experiment 4 (reject at all conventional levels) was due to: just more data which would increase the power; or of different data qualitatively; or of the different model, we did two checks. First, we estimated the same model as in Experiment 3 with the data of Experiment 4, and we reject one structure at all conventional levels (so this says it is not the model). Second, we kept the model as estimated in this section, and calculated Λ for the data of the experiment, but with 74 subjects, instead of 313: we kept the distributions and μ's as in (19), but allocated 74 subjects in all cells (instead of 313). Then we generated millions of samples of 74 subjects, and again the Λ for those samples was less than the estimated Λ in less than one tenth of 1% of the cases. So again, we rejected one signaling structure, indicating that the rejection is not due to the sample size. Therefore, we reject that the explanation for the "Unskilled and Unaware" pattern is due to regression to the mean, and that the explanation is that the data is different from that in Experiment 3.

In order to double check that the rejection is not due to the way we categorized the data (in four intervals of 25%), we also carried out the exercise of this section in a model in which (like in experiment 3) subjects declared "less than 50%", "between 50 and 80%" and "above 80%", and one signaling structure is again rejected.

2.3 Experiment 5, Unskilled and Unaware, estimate individual signal structures econometrically.

In experiment 5 we are interested in two questions. The first is to estimate econometrically the signalling structures of individuals to assess whether the unskilled have noisier information than the skilled; since the unskilled are wrong, by a larger margin, than the skilled, this could be interpreted as being tantamount to "the unskilled are unaware". This is similar, but goes one step further, than the questions asked in experiments 3 and 4, as in that case we only considered the question of whether they were different. This experiment is also of interest, as we analyze the question with a different methodology.

The second question we analyze is an "intensive margin" of the unskilled and unaware hypothesis: do people have noisier estimates of their performances when they do poorly?

In the experiment individuals are told, for each of 6 multiple choice questions, and for each of 6 open questions, how often a similar population answered each question. Then they are asked (in an incentive compatible way, and before observing the question) what is the chance they think they will answer it properly. Then they answer the question, and report their posterior belief of having answered correctly.

The basic model is as follows: given the prior α of answering a question correctly, and the posterior μ, we want to estimate the parameters β and γ : the probabilities of observing a signal "$c$orrect" given that the answer was indeed "Correct", and of observing the signal c, given that the anwer was "Wrong"

$$\beta = P(c \mid C) \text{ and } \gamma = P(c \mid W) \Rightarrow P(C \mid c) = \frac{\alpha\beta}{\alpha\beta + (1-\alpha)\gamma} \text{ and } P(C \mid w) = \frac{\alpha(1-\beta)}{\alpha(1-\beta) + (1-\alpha)(1-}$$

Then, if the answer reported a posterior larger than the prior, we assume he observed a c signal, and conversely if it is lower:

$$\mu = P(C \mid c) \text{ if } \mu \geq \alpha \text{ and } \mu = P(C \mid w) \text{ if } \mu < \alpha.$$

From these equations we can estimate two linear equations in β and γ :

$$\begin{aligned} 0 &= \beta\alpha(1-\mu) - \gamma\mu(1-\alpha) \text{ if } \mu \geq \alpha \\ \alpha - \mu &= \beta\alpha(1-\mu) - \gamma\mu(1-\alpha) \text{ if } \mu < \alpha. \end{aligned}$$

The same 313 subjects as in Experiment 4 participated. Based on the estimation of the β and γ for the skilled and the unskilled, we actually find no support for the Dunning Kruger Hypothesis. The skilled are not more precise than the unskilled in predicting their performance, nor are subjects better at detecting a correct answer as compared to an incorrect answers. This is not due to low power: the estimated coefficients are very close and sometimes even significantly different to one another, but there is no trend: in half of the cases it is the unskilled who are actually more precise.

Table 1: Estimation of the signaling model

	(1) y
beta unskilled	0.780***
	(82.90)
gamma unskilled	0.0999***
	(21.83)
d_beta	-0.0729**
	(-3.22)
d_gamma	-0.0307***
	(-5.00)
Observations	3756

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

We conclude that Skilled and Unskilled do not have the same signalling structures. However, the information processing of the skilled is not better than that of the unskilled: while γ is lower (more precise signals when the question is incorrect), their β is also lower, indicating less of an ability to recognize when they do well.

Work in progress: study intensive margin, and run a horse race between two models (overconfident and noise vs. unskilled and unaware).

3 Appendix 1. Instructions Experiment 1.

Instructions for Experiment 1

4 Appendix 2. Instructions for Experiments 2,4 & 5.

Instructions

Welcome! This is an experiment in decision-making. If you follow the instructions and make good decisions you will earn a substantial amount of money. The money you earn will be paid to you in CASH at the end of the experiment. The experiment has three parts, and there is a show up fee of 5 euro that you will earn regardless of your choices. The entire

experiment will take place on computer terminals. Please do not talk or communicate to each other in any way and turn off your phones now.

5 Preamble: Measuring your beliefs about the likelihood of events

In this experiment, you will be taking various trivia and logic quizzes. About half of your earnings will depend on how well you did in these quizzes, while the other half will depend on how accurately you evaluate your own performance. In particular, you will be asked the likelihood of certain events, with questions such as: "What are the chances that you gave the correct answer in the question you just answered?" or "What are the chances that you performed better than the median subject?".

Here we explain the procedure that will be used throughout the experiment to reward you for the accuracy of your self-assessment.

As an illustration, suppose that you are asked the following question: *Who is the current Prime Minister of the United Kingdom?* to which you answer *Theresa May.* You are then asked: *What are the chances that the answer you just gave was correct?.*

Your answer to this question will be measured in **chances**, which go from 0 (standing for: I am absolutely sure that I gave the wrong answer) to 100 (standing for: I am absolutely sure that I gave the right answer). So for example:

- 50 means that there are exactly equal chances that you were right or wrong;
- 33.3 (that is, one third of 100) means that you think you have a 1-over-3 chance to be correct, or, in other words, that you have the same chances to be correct as the chances to cast a 6-face die and draw a number smaller or equal to 2.
- 75 means that you have the same chances to be correct as the chance that a white ball is drawn from a bag with 75 white ball and 25 blue balls; and so on.

Review questions:

- What are the chances that you toss a fair coin and you get Tails?
- In a multiple choice question with 4 options, if you blindly pick one at random, what are the chances that it will be correct?

5.1 Incentives: How are you rewarded for reporting your chances accurately

We follow a special procedure to reward you for your self-assessment. This procedure is a bit complicated but the important thing to remember is that it is designed so that it is in your best interest to report your most accurate guess about your real chances. The procedure is as follows:

On the screen you can visualize a virtual bag. The bag is currently empty and will be filled with 100 blue and white balls. The exact composition of the virtual bag will be determined at the end of the experiment by a random device that will pick one of the following possibilities with equal likelihood: (0 white, 100 blue), (1 white, 99 blue), (2 white, 98 blue) ... (99 white, 1 blue), (100 white, 0 blue). There is a prize that you have the chance to win by either betting on your answer being correct or by betting on a white draw from the virtual bag. Whether you prefer to bet on your answer being correct or on the white draw from the virtual bag depends on how many white balls are in the bag. When there are 0 white balls, you probably prefer to bet on your answer being correct as in most situations you have at least some chances to be correct, no matter how small, while you will never draw a white ball from a bag that contains only blue ball. On the other extreme, when there are 100 white balls, you will probably prefer to bet on the virtual bag rather than on your answer, because it is guaranteed that you will win the prize from a bag with such a composition, whereas a grain of doubt may remain about the correctness of your answer. Somewhere in between 0 and 100 there is a number of white balls that makes you indifferent between betting on the corretness of your answer and betting on the virtual bag. We interpret this number as the chances that your answer is correct. In other words, if you are indifferent between betting on the bag with x white balls and betting on the correctness of your answer, it means you think you have exactly x/100 of having answered correctly.

So to incentivize you to be truthful, after you report your chances p to be correct in a question, your payment will be determined as follows:

- If, at the end of the experiment, in the virtual bag there are more than p white balls, you will bet on the virtual bag. That is, we will draw a random ball from the bag,

and, if the ball is white you will 4 euros, if not you will earn 0 euros.

- If instead in the virtual bag there are p white balls or less, you will bet on the correctness of your answer. That is, if your answer is correct you win 4 euros and if it is incorrect you win 0 euro.

Take some time to verify that it is indeed in your best interest to state your chances truthfully. Suppose that you believe you have 70/100 chances that your answer was correct. Then it means that you prefer to bet on you answer being correct rather than to bet on a white draw from the bag, if in the bag there are fewer than 70 white balls. Viceversa, if in the bag there are more than 70 white balls, you prefer to draw from the bag and hope in a white draw, which has more than 70/100 chance to happen. Being truthful ensures that you always get the better deal between the two options, given your beliefs.

We will use this procedure several times throughout the experiment so make sure you understand it, and please feel free to ask any question.

Part 1

Please read carefully and then answer the review questions.

In the first part of the experiment you are asked to answer 12 questions extracted from trivia or logic quizzes. You will receive payments for giving the right answers. You will also be asked to make a self-assessment of your performance (that is, for each question, to report how likely you think it is that your answer was correct) and will be paid for the accuracy of your self-assessment.

Assessment of your chances of a correct guess

You will report not one, but two guesses about how likely it is that a question was answered correctly. The first time, before seeing the actual question, and the second time after seeing and answering the question.

This is the timeline of events:

1. **Initial self-assessment** In the first screen you are asked to report the chances that you will answer each of the 12 questions correctly. At this stage you still don't know

the precise questions, but you are given the following information: (1) whether it is a multiple choice question or an open question; and (2) the percentage of people who gave a correct answer for this question in previous experiments. See a sample of the screen in Figure 1. Note that the minimum chance you can report is 0 for open questions and 25 for multiple-choice questions. The reason is that, with multiple-choice questions, in the worst case scenario, where you have no idea about the answer, you will pick one option at random and this will have a 25/100 chance to be correct (4 options in total, each with ex-ante equal chances to be correct).

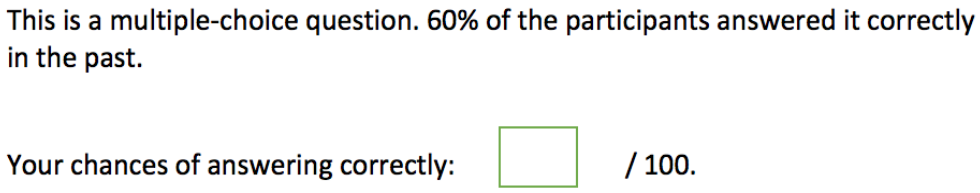


Figure 1. Initial Self −Assessment

2. **Answer to the questions**. Next you receive the questionnaire and submit your answers to the trivia and logic quiz. You don't have a formal time limit but, please, try to spend no more than 2 minutes per question.

3. **Posterior self-assessment** After each question, you are then asked to submit your best estimate of the chances that you answered that question correctly. Of course, you can now base your estimate upon knowing the actual question that was asked and the answer you gave. See a sample of the screen in Figure 2 for multiple-choice questions, and Figure 3 for open questions. For multiple-choice questions there is one and only one option which is correct and you always choose among 4 alternatives. In open questions you have to type in the answer. Your answer is not case sensitive and the software will handle small spelling mistakes.

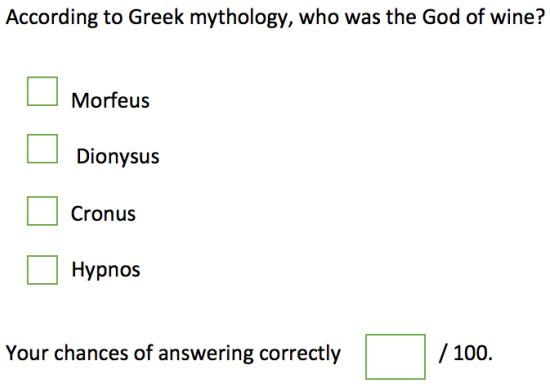


Figure 2. Multiple Choice Questions

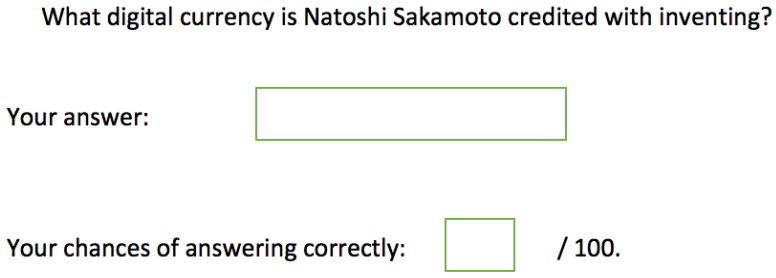


Figure 3. Open question.

Incentives and payoffs

You will be rewarded according to whether you answered the questions correctly and according to the accuracy of your self-assessment.

At the end of the experiment we will reveal the composition of the bag and extract a random question from the ones you answered. You will receive 4 euros if you gave the correct answer in that question and 0 euros otherwise.

Moreover, you will be rewarded for your self-assessment. A random draw will determine whether you are going to be paid according to your initial self-assessment or posterior self-assessment.

Next, your payment for self-assessment is determined as explained in the preamble. That is, the virtual bag will be filled with blue and white balls with the exact composition randomly determined.

The computer will then retrieve your reported chances **p** that your answer was correct in the selected question. Then your payment is determined as follows:

- If your stated chances **p** of a correct answer are larger than the number of white balls, you will be betting on your performance. That is you will receive 4 euros if your answer was correct and 0 euros if it was incorrect.

- If your stated chances **p** are lower than or equal to the number of white balls, then you will be paid according to a random draw from the virtual bag. In particular, you will receive 4 euros if we draw a white ball and 0 euros if we draw a blue ball.

Please feel free to ask any questions.

Review questions

1. Your stated chances are 53, in the virtual bag there are 20 white balls, we draw a white ball and your answer was correct.

2. Your stated chances are 35, in the virtual bag there are 80 white balls, we draw a blue ball and your answer was correct.

3. Your stated chances are 67, in the virtual bag there are 20 white balls, we draw a blue ball and your answer was incorrect.

4. Your stated chances are 15, in the virtual bag there are 80 white balls, we draw a white ball and your answer was incorrect.

When you are ready and comfortable with the instructions, click on the Next button to start part 1.

Part 2: Visual Task

In this part of the experiment you will perform 12 repetitions of the following exercise.

You will see a string of numbers blinking on the screen and will then have to type the numbers into the box appearing on the screen.

The duration of the blinks and number of elements in the string will vary across periods, hence remembering the string will be easier in some periods and more difficult in others.

You will do two practice rounds and then repeat this exercise 10 times for payment.

Payment:

At the end of the experiment one round will be selected at random. If, in that round, you reported the string of number correctly you will earn 2 euros, otherwise you will earn 0 euros.

Click on the Next button to proceed to the two sample rounds.

Part 3: Logic quiz

In this section you are asked to answer a logic Quiz. The Quiz consists of 12 multiple choice questions and you have 6 minutes to answer all the questions.

Self-assessment

Before you take the Quiz, we ask you to estimate how well you will do relative to the other subjects. Specifically, we ask you how likely you think it is that you will do better than half of the participants.

Here is how: After the quiz is complete, you will be assigned a ranking according to how many questions you answer correctly. The best performer among you will be assigned to rank 1, the second to rank 2 and so on.

We will then list the participants from highest rank to lowest rank and divide the subject pool into two equally sized-groups, an upper half and a lower half. For example, with 30 subjects the top 15 will ranked in the upper half and the other 15 will be ranked in the lower half. If two people are tied for 15th in terms of performance, then one of them will be randomly placed in the top half and one of them in the lower half.

We want you to tell us your best estimate of the probability that you are in the upper half. Your answer to this question will again be measured in **chances**, which go from 0 (standing for: I am absolutely sure that my score will not be in the upper half of the distribution) to 100 (standing for: I am absolutely sure that my score will be in the upper half of the distribution). So for example, 50 means that there are exactly equal chances that you score in the upper or the lower half and so on.

[Treatment 1: Payment based on lottery tickets and BDM]

Your payment for reporting your chances follows a procedure similar to the one outlined in the preamble, that is you will either bet on your placement in the upper half or on a white draw from the virtual bag. The only difference is that the prize now is 10 lottery tickets (each one worth a 3% chance of winning 20 euros). The procedure will go as follows. You will report your chance p of being in the upper half and then the computer will randomly determine the number of white balls in the virtual bag. Your payment will be determined as follows:

- If the number of white balls in the virtual bag is larger than p, then you will bet on the virtual bag. That is, a ball will be drawn from the virtual bag and if it is white you will receive the 10 lottery tickets (worth a total of 30% chance of winning 20 euros), otherwise you will get nothing.

- If instead the number of white balls is equal to or smaller than p, then you will be betting on your placement. That is, you will receive the 10 lottery tickets (worth a total of 30% chance of winning 20 euros) if your score indeed placed in the upper half of the distribution of scores, and otherwise you will get nothing.

[Treatment 2: Payment based on lottery tickets and VisualTask-BDM]

Your payment for reporting your chances follows a procedure similar to the one outlined in the preamble with two differences: (1) the prize for winning is now given by a number of lottery tickets (each one worth a 3% chance of winning 20 euros); and (2) your choice will not be between betting on your placement or betting on the virtual bag, but rather between betting on your placement in the Quiz or betting on your performance in the Visual Task.

The procedure is as follows: You will report your chances **p** of being in the upper half and then the computer will randomly determine the number of white balls in the virtual bag. Then:

- If the number of white balls is equal to or smaller than p, then you will be betting on your placement in the upper half. That is, you will receive the 10 lottery tickets (worth a total of 30% chance of winning 20 euros) if your score indeed placed in the upper half of the distribution, and otherwise you will get nothing.

- If instead the number of white balls in the virtual bag is larger than p, then you will bet on the visual task. That is, a ball will be drawn from the virtual bag and one of the 10 **10 or 12?** rounds that you completed in the visual task will be extracted at random (with each round having the exact same probability of being selected). If the ball is white and you were successful at the visual task in the extracted round, you will receive **N** lottery tickets (worth a total of **M** per cent chance of winning 20 euros) ; otherwise you receive zero euros. (Note: in the experiment, the numbers **N** and **M** were calibrated for each subject, depending on their success rate in the visual task).

Notice that given your success rate in the visual task, your chances of winning the lottery tickets after a white ball is drawn is P per cent.

The number of lottery tickets that you can win is calibrated on your performance in the visual task to ensure that it is indeed in your best interest to report the chances of being in the upper half accurately.

Before you state your chances of being in the upper half, you will answer 3 sample questions which are comparable in difficulty to the questions that you will find in the Quiz. There is no payment for the sample questions.

You are now ready to start the sample questions. Please click on the Next button now.

[The following is the message visualized on the screen after the sample questions]

What are your chances to be in the upper half of the scores distribution? Type a number between **0** (meaning: I have zero chance to be in the upper half) to **100** (meaning: I am absolutely sure I will be in the upper half of score distribution).

Reminder:

- The sample questions you just saw are of comparable difficulty to the actual questions you will encounter in the Quiz
- In past experiments, the better performing half of the subjects answered 7 or more questions correctly (out of 12) in the Quiz.

References

[1] Benoît, J.-P., Dubra, J. and Moore, D. A. (2015), Does the Better-than-average Effect show that People are Overconfident?: Two Experiments. Journal of the European Economic Association, 13: 293â€“329.

[2] Burks, Stephen, Jeffrey Carpenter, Lorenz Goette and Aldo Rustichini (2013), "Overconfidence and Social Signalling," Review of Economic Studies, 80(3), 949-83.

[3] Camerer, C., and Lovallo, D. (1999), Overconfidence and Excess Entry: An Experimental Approach. The American Economic Review 89.1 : 306â€“318.

[4] Clark, Jeremy and Lana Friesen (2009), "Overconfidence in Forecasts of Own Performance: An Experimental Study," Economic Journal, 119(1), 229-51.

[5] Ehrlinger, J., K. Johnson, M. Banner, D. Dunning and J. Kruger (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.

[6] Eil, David and Justin Rao (2011), "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," American Economic J: Microeconomics, 3, 114-38.

[7] Goodie, A. S. (2003) The effects of control on betting: paradoxical betting on items of high confidence with low value. J Exp Psychol Learn Mem Cogn: 29, 598-610.

[8] Goodie, Adam and Diana Young (2007), "The skill element in decision making under uncertainty: Control or competence?," Judgment and Decision Making, 2(3), pp. 189-203.

[9] Heath, Chip and Amos Tversky, (1991) "Preference and Belief: Ambiguity and Competence in Choice under Uncertainty," Journal of Risk and Uncertainty, 4, 5-28.

[10] Kruger, Justin and David Dunning (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**, 1121–34.

[11] Merkle, Christoph and Martin Weber (2011), "True Overconfidence: The Inability of Rational Information Processing to Account for Apparent Overconfidence," Organizational Behavior and Human Decision Processes, 116(2), 262-71.

[12] Moore, Don and Paul. J. Healy (2008), "The trouble with overconfidence," Psychological Review, 115(2), 502-517.