Supporting Information

Denrell and Liu 10.1073/pnas.1116048109

SI Text

How to Compute the Expected Skill. The performance of agent i in period t is $P_{i,t} = u_i + \varepsilon_{i,t}$, where u_i represents the average performance or skill of agent i and $\varepsilon_{i,t}$ is a noise term that is independent of u_i and has expected value zero.

As described in the main text, we assume that u_i is drawn from a normal distribution with mean zero and SD $\sigma_{i,u}$. Moreover, $\sigma_{i,u}$ is drawn from a gamma distribution with parameters $s, 1/s$ [i.e., the density of $\sigma_{i,u}$ is $(1/s)^{-s}\sigma_{i,u}^{s-1}e^{-\sigma_{i,u}/(1/s)}/\Gamma(s)$]. Similarly, $\varepsilon_{i,t}$ is drawn from a normal distribution with mean zero and SD $\sigma_{i,e}$, and $\sigma_{i,e}$ is drawn, independently of $\sigma_{i,u}$, from a gamma distribution with parameters $n, 1/n$.

To examine whether high performance is necessarily an indicator of high skill, we consider an observer who has observed the level of performance achieved by one agent in a particular period, $P_{i,t}$. The task of the observer is to estimate the agents' skill based only on information about the observed performance. Because our focus is on how much observers could, at best, learn from performance, we assume that the observer is rational and capable of implementing Bayes' rule to calculate the expected ability given the observed performance. That is, the observer computes $E[u_i \mid P_{i,t} = p_{i,t}]$, the posterior expected value of u_i given an observed performance of $P_{i,t} = p_{i,t}$.

To compute $E[u_i \mid P_{i,t} = p_{i,t}]$, note first that if $\sigma_{i,u}$ and $\sigma_{i,e}$ are known, then standard results (1) imply

$$E\left[u_i \mid P_{i,t} = p_{i,t}, \sigma_{i,u}, \sigma_{i,e}\right] = \frac{p_{i,t}\sigma_{i,u}^2}{\sigma_{i,u}^2 + \sigma_{i,e}^2}.$$

To get $E[u_i \mid P_{i,t} = p_{i,t}]$, we integrate over all values of $\sigma_{i,u}$ and $\sigma_{i,e}$:

$$E\left[u_i \mid P_{i,t} = p_{i,t}\right] = \int_{\sigma_{i,u}=0}^{\infty}\int_{\sigma_{i,u}=0}^{\infty} \frac{p_{i,t}\sigma_{i,u}^2}{\sigma_{i,u}^2 + \sigma_{i,e}^2} f\left(\sigma_{i,u}, \sigma_{i,e} \mid P_{i,t} = p_{i,t}\right) d\sigma_{i,u} d\sigma_{i,e}. \quad [S1]$$

Here, $f(\sigma_{i,u}, \sigma_{i,e} \mid P_{i,t} = p_{i,t})$ is the conditional joint density of $\sigma_{i,u}$ and $\sigma_{i,e}$ given the observed performance. To find $f(\sigma_{i,u}, \sigma_{i,e} \mid P_{i,t} = p_{i,t})$, we use Bayes' rule (and note that conditional on $\sigma_{i,u}$ and $\sigma_{i,e}$, performance is normally distributed with mean zero and variance $\sigma_{i,u}^2 + \sigma_{i,e}^2$). That is,

$$f\left(\sigma_{i,u}, \sigma_{i,e} | P_{i,t} = p_{i,t}\right) = \frac{\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\sigma_{i,u}^2 + \sigma_{i,e}^2}} e^{\frac{-p_{i,t}^2}{2\left(\sigma_{i,u}^2 + \sigma_{i,e}^2\right)}} g_u\left(\sigma_{i,u}\right) g_e\left(\sigma_{i,e}\right)}{\int_{\sigma_{i,u}=0}^{\infty}\int_{\sigma_{i,e}=0}^{\infty} \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\sigma_{i,u}^2 + \sigma_{i,e}^2}} e^{\frac{-p_{i,t}^2}{2\left(\sigma_{i,u}^2 + \sigma_{i,e}^2\right)}} g_u\left(\sigma_{i,u}\right) g_e\left(\sigma_{i,e}\right) d\sigma_{i,u} d\sigma_{i,e}}, \quad [S2]$$

where $g_u(\sigma_{i,u})$ is the density of $\sigma_{i,u}$ and $g_e(\sigma_{i,e})$ is the density of $\sigma_{i,e}$.

Based on Eqs. **S1** and **S2**, we can find $E[u_i \mid P_{i,t} = p_{i,t}]$ for any combination of (n, s) by numerically computing the associated integrals. Fig. 3 provides an illustration and shows how a nonmonotonic pattern between performance and skill emerges when n is smaller than s ($n = 1$ and $s = 5$). Extensive checks show that whenever $n < s$, there is a value p^* such that $E[u_i \mid P_{i,t} = p_{i,t}]$ decreases in $p_{i,t}$ for $p_{i,t} > p^*$ (the smaller the distance between n and s, the higher the value of p^*).

Our model makes a number of assumptions about the skill and noise distributions and about what can be observed. Nevertheless, the basic result does not hinge upon these assumptions. Extensive computations and simulations show that the same basic result—that inferences about ability from observed performance can be nonmonotonic—holds more generally.

Consider first the number of observations the observer has access to. In the above model, we assumed that the observer could only observe one performance for each individual and had to make inferences about the underlying ability based on this single observation. In reality, observers may be able to observe several performances for each individual. Does the basic result still hold in this case? Overall, the answer is yes. For example, consider a special case of the above model (with no heterogeneity in the skill distribution, $s \to \infty$, and heterogeneity in the noise distribution, $n = 1$) and suppose that the observer had access to two observations rather than one. That is, the observer has access to two pieces of information, $P_{i,1} = p_{i,1}$ and $P_{i,2} = p_{i,2}$, and wants to compute $E[u_i \mid P_{i,1} = p_{i,1}, P_{i,2} = p_{i,2}]$. We assume that $P_{i,1} = p_{i,1}$ and $P_{i,2} = p_{i,2}$ are independent draws from the same performance distribution. In this case the basic result still holds, for the cases we have computed. That is, $E[u_i \mid P_{i,1} = p_{i,1}, P_{i,2} = p_{i,2}]$ is at a maximum for intermediary levels of performance. Computations show that the same result holds even if the observer has access to more observations. The magnitude of the effect, however, declines when the observer gets access to more observations, because access to more observations implies that the observers learn the underlying skill with higher precision.

Consider next alternative assumptions about the skill and noise distributions. We assumed that the skill and noise terms were drawn from a normal distribution with SDs drawn from gamma distributions. We have tried other distributions for the skill and noise terms and get similar results. For example, the SDs could be drawn from exponential distributions [with density $(1/b)e^{-x/b}$]. Computations show that expected skill is decreasing in observed performance for extreme performances whenever $\sigma_{i,e}$ has a higher variance than $\sigma_{i,u}$ (i.e., higher b). No general sufficient condition seems to be known in the literature, however, regarding what type of skill and noise distributions our result holds for.

What can be demonstrated is that if the skill and noise distributions are identical, then $E[u_i \mid P_{i,t} = p_{i,t}]$ is increasing in $p_{i,t}$. To show this, note that $P_{i,t} = u_i + \varepsilon_{i,t}$. Taking expectations on both sides, we get $p_{i,t} = E[u_i \mid P_{i,t} = p_{i,t}] + E[e_{i,t} \mid P_{i,t} = p_{i,t}]$. By symmetry it follows that $E[u_i \mid P_{i,t} = p_{i,t}] = E[e_{i,t} \mid P_{i,t} = p_{i,t}]$. Denote their common value by c. It follows that $p_{i,t} = 2c$, or $c = 0.5p_{i,t}$.

Also, a necessary condition for our result is well-known: The conditional density $h(p_{i,t} \mid u_i)$ must violate the monotone likeli-

hood ratio property [MLRP (2)]. The MLRP requires that for $u_2 > u_1$, $h(p_{i,t} \mid u_2)/h(p_{i,t} \mid u_1)$ is increasing in $p_{i,t}$. When $P_{i,t} = u_i + \varepsilon_{i,t}$, $h(p_{i,t} \mid u_i) = f(p_{i,t} - u_i)$, where f is the density of the noise term. Thus, a violation of the MLRP requires that $f(p_{i,t} - u_i)$ violates the MLRP. If there is a violation, there exists a prior distribution $g(u_i)$ such that $E[u_i \mid P_{i,t} = p_{i,t}]$ is not increasing in $p_{i,t}$. It follows that our result cannot happen when the noise term is normally distributed (for which $f(p_{i,t} - u_i)$ satisfies the MLRP) and our result could happen for some fat-tailed noise distributions, such as the Cauchy or the t distribution, which violate the MLRP.

If the support of the skill distribution is bounded, a recent paper (3) shows that reversals of conditional expectations will occur if the noise distribution has much wider support, that is, has fatter tails. For example, the skill distribution might range from −5 to 5, whereas the noise distribution might range from −30 to 30. In such a case, an extreme performance cannot be due to very high average skill (because it is bounded) but has to be due to noise. In reality, observers can seldom exclude the possibility of very high skill. Nevertheless, this result regarding skill distributions with bounded support does lend support to our conjecture that reversals of conditional expectations occur when the noise term is more fat-tailed and thus more likely than the talent term to generate extreme values.

Experimental Design. The purpose of the experiment was to examine whether people were able to learn a nonmonotonic association between performance and skill, given that past studies show that people tend to assume that higher performers are more skilled.

Data. We examined people's inference using four datasets with different patterns between performance and skill. In two datasets (datasets 1 and 3), the association between performance and skill is nonmonotonic (the highest performers do not have the highest level of skill), and in two datasets (datasets 2 and 4), the association between performance and skill is monotonic (higher performance indicates higher expected skill).

All datasets were generated by the model described in the above section in which $P_{i,t} = u_i + \varepsilon_{i,t}$. The datasets differ only in the parameters s and n. In dataset 1, there is no heterogeneity in skill distribution but heterogeneity in noise distribution. That is, u_i is drawn from a normal distribution with mean zero and SD equal to one, and $\sigma_{i,e}$ is drawn from a gamma distribution with parameters $1, 1$ ($n = 1$). For dataset 2, $s = 1$ and $n = 5$, which implies that the association between performance and ability is monotonic. For dataset 3, $s = 5$ and $n = 1$, which implies that the association between performance and ability is nonmonotonic. For dataset 4, $s = 1$ and $n = 1$, which implies that the association between performance and ability is monotonic.

To ensure that participants in the experiment observed high, intermediary, and low levels of performance, we did stratified sampling rather than random sampling from the above models. That is, to construct each dataset, we first simulated a large number of performance–skill pairs following the above models. Then we divided all of the simulated data into 10 groups according to their performance levels, ranging from low to high observed performance. We then randomly drew 20 observations from each of the 10 groups to get the 200 observations we needed for the experiment. We calculated the average performance and skill levels for each of the 10 groups to ensure that the association between performance and ability of the sampled 200 agents followed the pattern we describe above.

Experimental procedure. Participants were asked to predict the future performance (i.e., average skill) of a sales representative based on data on this sales representative's past performance. After a participant made a prediction, the actual future performance was displayed. This was repeated for 200 different sales representatives, resulting in 200 predictions per participant. The instructions emphasized that the task was to learn the association between past and future performance from the data displayed rather than relying on prior knowledge about performance distributions in sales.

The reward a participant obtained depended on how accurate his or her predictions were. Participants started with a reward of 20 pounds, and for each prediction a penalty was deducted from this sum based on the discrepancy between the predicted and the actual future performance.

Participants. Participants were recruited from a behavioral laboratory in the United Kingdom (32 responses, of which 18 were male and 14 were female; average age was 27.21 y with an SD of 6.36) and from Amazon Mechanical Turk (181 responses, of which 99 were male and 56 were female; 26 were not willing to respond to the question on sex; average age was 29.07 y with an SD of 9.07). Sixty-three participants received dataset 1, 43 participants received dataset 2, 56 participants received dataset 3, and 51 participants received dataset 4.

Analysis and results. For each participant, we first grouped the 200 predictions into 10 groups based on the observed performance (ranging from low to high observed performance). We then calculated the average prediction for each group of observations.

We were interested in whether participants' predictions were monotonically increasing: Are their predictions highest for the top group, lower for the second group, lower still for the third group, and so forth? We classified a participant as making "nonmonotonic" predictions if (*i*) the average prediction for the first group is lower than the average prediction for the second group or (*ii*) the average prediction for the first group is lower than the average prediction for the third group or (*iii*) the average prediction for the second group is lower than the average prediction for the third group. Note that our focus is on possible nonmonotonicity in predictions for relatively high performers rather than on predictions for relatively low performers.

The result shows that few participants can be classified as responding according to a nonmonotonic pattern even if the underlying association is nonmonotonic and participants had ample time, precise feedback, and incentives to be accurate. Among the 119 participants who received dataset 1 or dataset 3 (in which there was a nonmonotonic association between performance and expected skill), 69 (58%) did not display a nonmonotonic pattern in their responses. Among the 94 participants who received dataset 2 or dataset 4 (in which the association between performance and expected skill was monotonic), 80 (86%) did not display a nonmonotonic pattern. These results suggest that it is a challenge for people to switch to a nonlinear model when evaluating performance, despite the presence of immediate and clear feedback.

Simulation. The purpose of the simulation was to compare the predictive accuracy of a linear and a third-degree polynomial regression model. In the simulation, each model was fitted to data on individual performance and skill levels. The estimated models were then used to predict the skill level of a new individual based on this individual's observed performance.

The simulation was constructed as follows. In every period, the performance, p_i, of an individual with an unknown level of skill, u_i, is observed. The task is to predict the level of skill, u_i, based on the observed performance. After the prediction, the skill level is revealed. Predictions are based on data on skill levels and observed levels of performance acquired over time. After each period, another performance–skill pair is observed. After t periods, data on t performance–skill pairs are available.

The linear model assumes that $u_i = a + bp_i$ and estimates a and b by ordinary least square regression using past data. The predicted level of skill for the performance level observed in period t is $\hat{u}_t = \hat{a} + \hat{b}p_t$. The third-degree polynomial model assumes that $u_i = a + bp_i + cp_i^2 + dp_i^3$ and estimates a, b, c, and

d by minimizing the sum of squared deviation using past data. The predicted level of skill for the performance level observed in period t is $\hat{u}_t = \hat{a} + \hat{b}p_t + \hat{c}p_t^2 + \hat{d}p_t^3$.

The association between performance and skill was assumed to be nonmonotonic. In particular, the performance level in period t is constructed as follows: $p_t = u_t + \varepsilon_t$, where u_t is drawn from a normal distribution with mean zero and SD equal to $\sigma_{t,u}$ and $\sigma_{t,u}$ is drawn from a gamma distribution with parameters $(5, 1/5)$, and ε_t is drawn, independently of u_i, from a normal distribution with mean zero and SD equal to $\sigma_{t,e}$ and $\sigma_{t,e}$ is drawn from a gamma distribution with parameters $(1, 1)$. This specification implies that the noise distribution is more fat-tailed than the skill distribution. It follows that there is a nonmonotonic association between performance and skill, that is, the highest performers are not the most skilled (Fig. 3).

In every period, we examine whether the prediction of the linear or the polynomial model is more accurate, that is, closer to the actual skill level. We repeated this simulation 1 million times and computed, for each period, the percentage of the simulations in which the linear model made the most accurate prediction. Although the polynomial model can better fit the true nonmonotonic association between p_i and u_i, Fig. S1 shows that the linear model is nevertheless more accurate for the first 20 predictions, namely when the sample size is small.

1. DeGroot M (1970) *Optimal Statistical Decisions* (McGraw-Hill, New York, NY).
2. Milgrom P (1981) Goods news and bad news: Representation theorems and applications. *Bell J Econ* 12(2):380–391.
3. Chambers CP, Healy PJ (2011) Reversals of signal-posterior monotonicity for any bounded prior. *Math Soc Sci* 61(3):178–180.

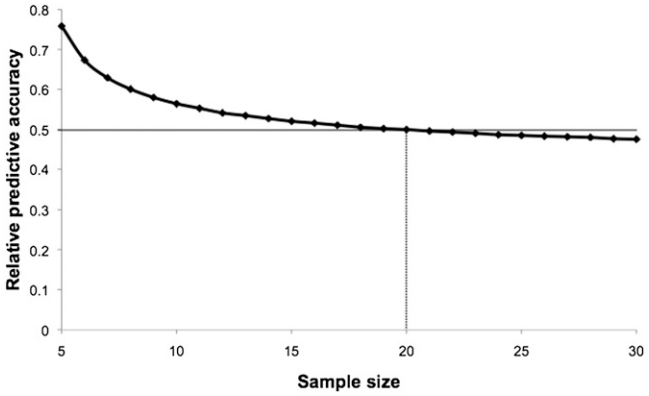

Fig. S1. Illustration of how a linear model can outperform a third-degree polynomial model (which better fits the assumed nonmonotonic relationship between performance and skill) in a sequential prediction task when information is scarce. Relative predictive accuracy is the proportion of 1 million simulations in which the predicted skill level from a linear model is closer to the true value than the prediction from a third-degree polynomial model.