Incentives and timing in relative performance judgments.

A field experiment

Michal Krawczyk

September 13, 2010

Abstract:

Several studies have identified the "better than average" effect – the tendency of most people to think they are better than most other people on most dimensions. The effect would have profound consequences (see e.g. Barber and Odean (2001)). These findings are predominantly based on non-incentivized, non-verifiable self-reports. The current study looks at the impact of incentives to judge one's abilities accurately in a framed field experiment. Nearly 400 students were asked to predict whether they would do better or worse than average in an exam. The most important findings are that subjects tend

to show more confidence when incentivized and when asked before the exam rather than afterwards. The first effect shows up particularly in females.

Keywords: better-than-average effect, overconfidence, underconfidence, incentives, gender differences

JEL classification: C93, D04, D84.

1 Introduction

Studies in psychology show that most individuals tend to be overly confident about own chance of success in most enterprises. In particular, the "better than average" effect (BTAE) (Svenson 1981) has been identified, under which most people in a given population think they are better on a given dimension than most other people in the same population, which is clearly a logical impossibility.[1] Such features may have massive economic consequences (see for example, Malmendier and Tate (2005), Odean (1998) and the discussion in Clark and Friesen (2009)). To be sure, subsequent research showed numerous factors mitigating the effect. For example, comparisons against a specific other tend to be less favorable than against an abstract one (Alicke, Klotz, Breitenbecher, Yurak, and Vredenburg 1995). There are also important cultural differences (cf. Endo, Heine, and Lehman (2000))

[1]BTAE is a special case of what Larrick, Burson, and Soll (2007) and Moore and Healy (2008) call overplacement – the tendency to place oneself too high in a ranking.

Most of the studies in this research tradition involve comparisons that are unlikely to be verified. For example, everyone can safely claim to be more "reliable" than an average person as long as this is cheap-talk only (and thus probably will, at a job interview or the first date in any case).

Therefore, as aptly argued by Hoelzl and Rustichini (2005) the better-than-average effect should be subject to further research involving feedback and *incentive* to judge one's abilities accurately. To the extent that deviation from correct predictions results from insufficient mental effort or the willingness to "look good", incentives may remove the bias. This will not be the case however, if observed overconfidence merely reflects genuine (if incorrect) beliefs.

The importance of feedback has long been recognized in the closely related literature on unrealistic optimism. An influential paper (Shepperd, Ouellette, and Fernandez 1996) claims that individuals often abandon excessive optimism as the "moment of truth" approaches. More recently, Van Dijk, Zeelenberg, and Van Der Pligt (2003) interpreted this tendency as a disappointment-avoidance strategy, arguing that two feats of feedback are instrumental: self-relevance and proximity in time. Whenever predictions and actual outcome are public, the anticipation of *shame* can lead to identical behavior – subjects claim that they will excel in the task while feedback is temporally distant. However, as it draws near, painful discrep-

ancy between publicly expressed high expectations and possible low outcome becomes a threat. Thus, subjects revert to more modest predictions. This self-restraint, however, can be overcome if promise of a reward dwarves the concerns for social image.

Unfortunately, the problem of prediction *incentivization* found less appeal, at least before the experimental revolution in economics. Yates, Lee, and Bush (1997) compared reported probability judgments in the context of general knowledge questions to judgments inferred from subjects' BDM-based evaluation of a gamble paying conditional on their answer to any given question being correct. Unexpectedly, they find that their American sample, overconfident in reported judgments became *more* overconfident in terms of their willingness to bet on own knowledge; no impact was identified in the Chinese sample (which, in general, was even more overconfident). Unfortunately, the elicitation method the author used was incentive-compatible under assumption of risk neutrality only.

More recently a few papers took on the issue. In a breakthrough contribution Hoelzl and Rustichini (2005) gave their laboratory subjects a choice between two conditions: "performance test" and "lottery". Subjects were expected to vote for the former if they believed they would perform above median in the test. Hoelzl and Rustichini (2005) find that subjects are over-

confident only when the task is easy,[2] whereas they switch to underconfidence if the task is difficult *and* success is rewarded with money.

The payment is thus conditional upon "success" (in a task or a lottery) not successful *prediction* of own performance. Subjects can thus vote for "performance test" condition because they think that it is fair if the payment is contingent on performance rather than luck (Krawczyk 2009). This could explain why more subjects in the study choose the task than declare belief in success in it. Some subjects could prefer "lottery" as, contrary to "task", it involved success determination on individual level, generally resulting in less ex-post inequality (Kroll and Davidovitz 2003). Finally, as the authors recognize ambiguity aversion could guide the choice.

Grieco and Hogarth (2009) also use trivia tests in a laboratory, letting subjects chose between a random and performance-contingent payments. Their study is more comprehensive, manipulating i.a. incentives and the mode of elicitation. Because, unlike in (Hoelzl and Rustichini 2005) the choices were generally made after the task and on individual level, thus avoiding the problems discussed above, except ambiguity aversion. The main finding of Grieco and Hogarth (2009) is that while over(under)confidence regarding the absolute score shows up for hard(easy) tasks as expected, there is no bias in placement (i.e. relative) judgments. Of special importance to my

[2]Such hard-easy discrepancy has been noted before, see e.g. Kruger (1999)

study is these authors' experiment two, where confidence before and after the task is directly compared within-subjects, finding no statistically significant difference.

Clark and Friesen (2009) also investigated overconfidence in a lab experiment involving real effort tasks (verbal decoding and function maximization). The result was generally negative, i.e. subjects were not biased in their predictions, particularly with respect to relative performance, no matter whether the incentives to predict correctly were present or not.

Another interesting attempt is due to Park and Santos-Pinto (2010). These authors elicit and incentivize predictions of relative performance of student poker tournament players. Overall, they find some overconfidence. The elicitation method assumes risk neutrality. It cannot be ruled out that individuals who are relatively over-confident and risk-seeking and overweight low probabilities of success actually self-select to amateur poker tournaments. This conjecture is corroborated by the fact that the bias was smaller in the chess tournament the authors also used in the study.

Finally, a closely related field study was run by Krajc (2008).The author asked university students to predict their own scores, others' mean scores and own relative standing (percentile) in two exams. He observed sizable overconfidence and overplacement that seemed to be decreasing over time (i.e. as students acquired more information). The author had only a small

sample of app. 50 students available.

To summarize, proper incentivization of predictions on own performance turns out to be quite challenging. Authors of but a few studies were able to develop clever methods to partly overcome the difficulties. The results obtained are mixed – the claim of universal overconfidence is not supported. To the best of my knowledge, my study is the first one to directly test the role of incentives and timing in a field experiment on the BTAE.

Indeed, the current study looks at the impact of incentives not in a laboratory but rather in subjects' natural environment. In a framed field experiment, university students were asked to predict whether they did better or worse than the mean. Subjects had to give the answer before the exam in one case and after the exam in another; half the subjects were incentivized to give the correct answer. We see that subjects are not universally overconfident. The BTAE tends to show a) when monetary prize is promised, b) if the assessment takes place *before* the exam and c) in males.

2 Design

The experiment was run during two final exams in Microeconomics at the Faculty of Economics, University of Warsaw in February ("Micro 103") and June 2010 ("Micro 102"). No subject took part in both of them. Each

student taking an exam was asked to predict whether their score (in points) would be higher or lower than the mean of all students' scores. They would do so in one of two treatments: Money (MT) and No-Money (nMT). In the former students were informed that one correct prediction would be picked at random and rewarded with PLN250, equivalent of app. 20 hours of a basic campus job. The instructions under MT read "Please try to predict whether your score [in points] from today's exam will be higher or lower than the average score (arithmetic mean of all the scores). A prize of PLN250 will be awarded to one of the students who give the correct answer. The winner will be randomly selected on [date, time, location]. The name of the winner (but not the score) will be announced and (s)he will be notified by e-mail. Needless to say, your response to this question will have no impact on your grade.

Now, do you think your score will be (please underscore):

higher than average OR lower than average.

Thanks a lot!"

The procedure was later implemented as promised. The nMT questionnaire was identical, except that no monetary prize was mentioned or paid out for that matter.

While in the experiment, subjects were unaware of the existence of the other treatment, as all students in the same lecture hall would be assigned

to the same treatment. Assignment to lecture halls, in turn, was random. The important (and only) difference between the two sessions was that the experiment was run directly *before* Micro 102 exam (condition PRE) and directly *after* Micro 103 (POST).

From the viewpoint of interest in the dynamics of overconfidence, it would have been tempting to run a within-subject comparison, asking the subjects to predict their success both before and after the exam. However, such a design would lead to certain difficulties, especially in the Money Treatment. Standard concerns in within-subject experiments (such as experimenter demand effect and consistency-seeking) aside, subjects could try to "hedge" by predicting success before the exam and failure thereafter or vice versa, largely regardless of their true beliefs.[3]

The current design design distinguishes itself from previous studies, which helps to avoid some of the problems mentioned yet may lead to new ones. Crucially, it involves rewards for correct predictions rather than choice between betting on own success or betting on dice as in (Hoelzl and Rustichini 2005), (Grieco and Hogarth 2009). One could thus wonder whether monetary incentives to predict correctly may distort the preference to perform well in the exam in the case of Micro 103 (i.e., when experiment is conducted before

[3]This could be prevented by means of a mild form of deception, whereby students would be unexpectedly asked to "update" their predictions after the exam.

the exam).[4] Indeed, it is easy to fulfill the prediction that one's score will be worse than average, rendering the question of prediction accuracy futile. However (and this is one of the crucial benefits of the field study), students clearly have strong motivation to pass the exam that overshadows the relatively unlikely event of winning the prize; indeed, my results show that the score did not differ between experimental treatments. As excelling in an exam is clearly a legitimate source of benefits, there is no risk of perceiving superior performance as anti-social either (as could be the case in laboratory studies using highly abstract and mundane tasks). Finally, the exams were compulsory for the students majoring in economics, so the question of selection appears less pressing than in (Park and Santos-Pinto 2010).

However, a careful reader may note that the lottery scheme used in the MT may not, strictly speaking, induce a student to choose the option ("better" or "worse") she perceives as most probable. This is because being better than average may be correlated with the number of people being correct and thus the odds of being selected. For instance, consider a student who thinks her absolute score is, say, around 20 points. She is not sure whether this is below or above group average. For the sake of argument, let us assume she thinks that two states of the world are possible: the questions, from the

[4]This is a concern in (Clark and Friesen 2009) who refer to it as the *moral hazard* problem.

perspective of other students, may be "standard" or "non-standard". In the former case, the scores are generally high and easy to predict. Thus, her own score is worse than average and most, say, all students correctly predict their standing. If, on the other hand, questions are non-standard, her score of 20 is above average and confused students only guess their relative score correctly half of the time. Now, if she believes that the probability of the questions being standard (and thus her score being below average) is 60%, it still pays better to bet on her score being above average, because this increases her probability of obtaining the reward from $\frac{.6}{N}$ to app. $\frac{.8}{N/2} = \frac{.8}{N}$, where N stands for the number of students (she thinks are in the MT). On the other hand, if she knows the literature on overconfidence, specifically the hard-easy effect, she may be tempted to reformulate the problem slightly, expecting the questions to be "easy" or "hard" for other students. Correspondingly, her score will be above average in the former case only. She may think that other students will judge their performance in a more biased (overly confident) way if questions are easy. As BTAE, other things being equal, lowers prediction accuracy, in expectation more people will answer correctly and thus the odds of getting the correct guess rewarded conditional on the questions being hard and thus her score of 20 – better than average, will be lower. She may thus end up betting on this outcome even if she thinks it is less likely than the other. However, as the reasoning of the type sketched

above is very demanding (in terms of the number of steps and background knowledge required), I find it highly implausible that more than a negligable fraction of students ever considered it within the limited time they had to predict their performance.[5] It seems obvious to me and is in line with the informal discussions we had after the experiments that most student were only concerned with the probability of being above or below average, not the conditional expected value of the reward in either case.

If that conjecture is correct, then, assuming that subjects are able to judge their performance in a bias-free way, half of them should expect to score above average[6].

Basing on the literature discussed in the preceding section, the following hypotheses may be put forward. First, subjects are overall overconfident. Second, they report less confidence as the feedback approaches (i.e. under POST)[7] for it brings with it the possibility of a shameful disappointment. Third, monetary incentives can hamper the latter effect. Fourth, money may reduce excessive confidence in the PRE condition; finally, male subjects tend to be more overconfident.

[5]Even if some did, they could still split in the two opposing camps described above.

[6]This need not be true if the distribution is skewed. Here, the difference between the mean and the median was less than .5 points (out of 30 or 50 in Micro 103, 102 resp.).

[7]Students typically engage in lengthy discussions with their peers aimed at finding the correct solutions directly after the exam (and, in the case of exams in questions, actually receive their scores within 48 hours). Thus the moment of truth was indeed nigh in the case of Micro 103.

3 Results

Two hundred twenty students participated in the Micro 102 exam (PRE), of which 201 answered the focal question (response rate was slightly higher in the Money treatment and did not correlate with gender or score on the test). Table 1 shows the essential results – subjects were overconfident in the Money Treatment only ($p < .01$, n.s. in nMT). Further, male subjects appeared more overconfident than females (specifically in the nMT).

Table 1: *Frequencies of expecting a higher-than-average score in Micro 102 (PRE), no. of obs. in parentheses*

treatment	females	males	total
nMT	.46	.65	.57
	(37)	(46)	(83)
MT	.80	.77	.79
	(56)	(62)	(118)

It seems interesting to find out how the monetary incentive affected the correctness of predictions. It turns out that 72% of subjects predicted correctly in the nMT and only 63% in the MT. This difference, however, was not significant.

The results for the Micro 103 exam were similar, with a general shift towards under-confidence. Of 179 students present, 160 answered the question (again, the rate was higher in the Money Treatment and did not correlate with gender or exam score).

Table 2 shows that subjects were underconfident in the nMT ($p < .01$) and overconfident in the MT ($p = .01$). Again, male subjects are somewhat more overconfident, in this case in both treatments. Again, there was no significant difference in the rate of correct predictions, 64% in nMT and 71% in MT.

Table 2: *Frequencies of expecting a higher-than-average score in Micro 103 (POST), no. of obs. in parentheses*

treatment	females	males	total
nMT	.30	.41	.34
	(52)	(32)	(84)
MT	.61	.68	.64
	(46)	(30)	(76)

The main treatment effect may not be explained by selection – the differneces between MT and nMT would remain significant even if all nMT non-responders had declared belief in being better than average under while MT non-responders – the reverse. This would have been highly implausiable, we may add, given that choosing to respond did not correlate with confidence-related variables such as score or gender under either treatment.

Figure 3 displays the locally-weighed smoothed scatter plot of predictions against actual relative score (score divided by maximum possible score) in all four treatments. One can see the expected positive relationship between the two variables, the strength of which does not seem to differ across treat-

ments. The graphs are more elevated under MT and under PRE, as discussed before.[8]

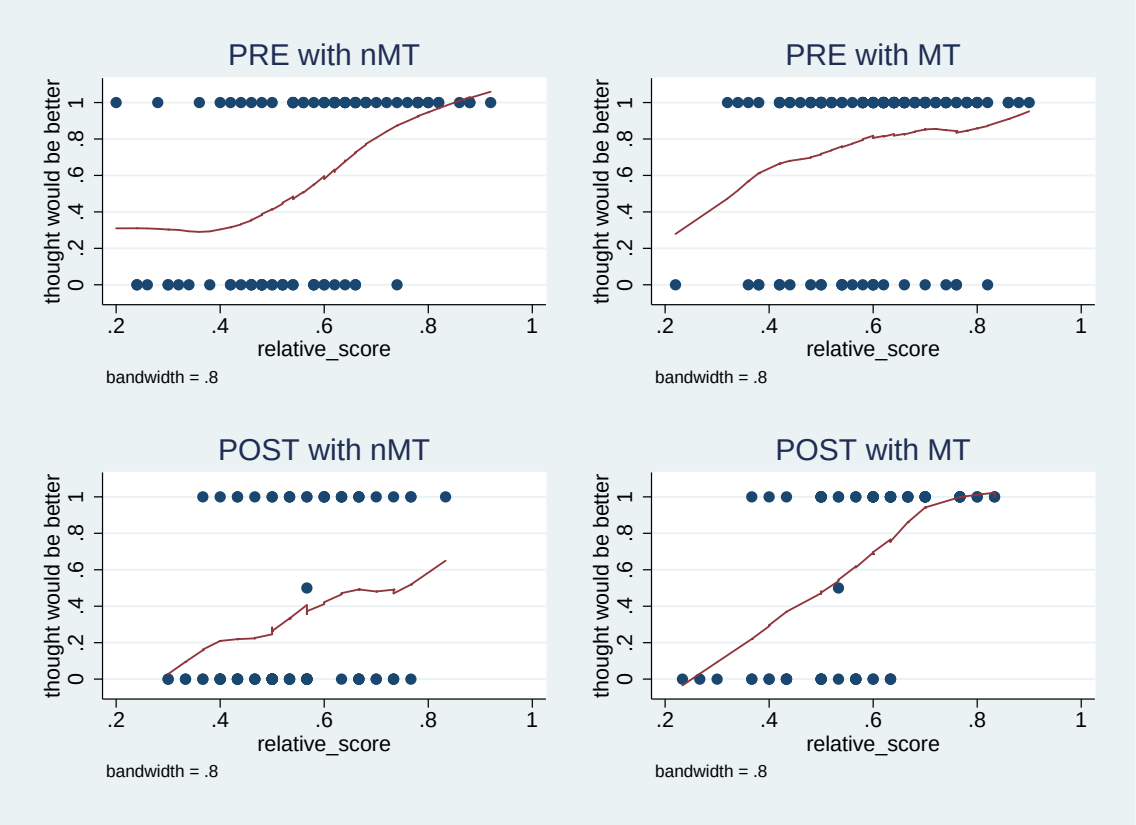

Figure 1: Locally weighted scatter plot smoothing of BTA against relative score, by treatment

Finally, these findings may be confirmed with a probit regression, see Table 3. Monetary incentives (MT) lead to more optimism while proximity of feedback (POST) to pessimism (the interaction term turns out to be non-significant). Not surprisingly, students actually scoring higher (rel score)

[8]The dots in the very center of the POST-nMT and POST-MT figures represent two individuals who wrote down the highly implausible prediction of their score being exactly equal to the average and were, in fact, remarkably close to being correct.

predict BTA scores more often (although the score being truly better than average, tBTA, has no additional impact) and males are more confident but only in the absence of monetary incentives. This marginally significant interaction between monetary incentive and gender is consistent with the findings of differences in self-presentation style (Gould and Slone 1982), (Daubman, Heatherington, and Ahn 1992) – females may be more inclined to underreport their expectations regarding own performance, as long as rewards for correct predictions are absent.

Table 3: *Determinants of BTA belief: a probit regression*

BTA	Coef.	Std. Err.	$P > \|z\|$
MT	.843	.202	.000
POST	$-.415$	.152	.006
tBTA	.101	.244	.679
rel score	3.710	.960	.000
male	.483	.213	.023
male*MT	$-.490$	.296	.099
cons	$. - 2.234$	.493	.000
N	361		
Pseudo R^2	.205		

4 Conclusion

This is the first field experiment to systematically investigate the impact of incentives and timing on the BTAE. The conjecture that the effect will go away when subjects are incentivized to think twice is rejected. As a matter

of fact, participants predict more optimistically in the Money Treatment.

This result may seem partly at odds with the findings of Hoelzl and Rustichini (2005). However, these authors only observe that incentives lead to less confidence in a difficult, unfamiliar task, while the tasks used in the current study was well known to subjects and not very hard. Further, if subjects employ a reasonable heuristic "don't take a hard task", it will lead to underconfidence in the choosing paradigm of Hoelzl and Rustichini but not the predicting paradigm used here.

The most appealing combination of factors to explain my data seems to be the following: First, subjects were generally overconfident; second, however, they considered it inappropriate to display the overly positive view of oneself (this is true in particular for female students), especially as the possibility of inglorious negative verification draws near. Third, this (feminine) modesty was overridden by the promise of a reward in the Money Treatment. Previous literature suggest that this finding may be culture-specific (Cai, Brown, Deng, and Oakes 2007). Overall, while monetary incentives are found to make a difference in the experiment, there is no evidence that the BTAE arises due to insufficient motivation to predict correctly.

University of Warsaw

References

Alicke, M., M. Klotz, D. Breitenbecher, T. Yurak, and D. Vredenburg (1995): "Personal contact, individuation, and the better-than-average effect.," *Journal of Personality and Social Psychology*, 68(5), 804–825.

Barber, B., and T. Odean (2001): "Boys Will be Boys: Gender, Overconfidence, and Common Stock Investment*," *Quarterly Journal of Economics*, 116(1), 261–292.

Cai, H., J. Brown, C. Deng, and M. Oakes (2007): "Self-esteem and culture: Differences in cognitive self-evaluations or affective self-regard?," *Asian Journal of Social Psychology*, 10(3), 162–170.

Clark, J., and L. Friesen (2009): "Overconfidence in Forecasts of Own Performance: An Experimental Study*," *The Economic Journal*, 119(534), 229–251.

Daubman, K., L. Heatherington, and A. Ahn (1992): "Gender and the self-presentation of academic achievement," *Sex Roles*, 27(3), 187–204.

Endo, Y., S. Heine, and D. Lehman (2000): "Culture and positive illusions in close relationships: How my relationships are better than yours," *Personality and Social Psychology Bulletin*, 26(12), 1571.

Gould, R., and C. Slone (1982): "The" feminine modesty" effect: A self-presentational interpretation of sex differences in causal attribution," *Personality and Social Psychology Bulletin*, 8(3), 477.

Grieco, D., and R. Hogarth (2009): "Overconfidence in absolute and relative performance: The regression hypothesis and Bayesian updating," *Journal of Economic Psychology*, 30(5), 756–771.

Hoelzl, E., and A. Rustichini (2005): "Overconfident: Do You Put Your Money On It?*," *The Economic Journal*, 115(503), 305–318.

Krajc, M. (2008): "Are the Unskilled Really That Unaware? Understanding Seemingly Biased Self-Assessments," *CERGE-EI Working Papers*.

Krawczyk, M. (2009): "A glimpse through the veil of ignorance: equality of opportunity and support for redistribution," *Journal of Public Economics*, 94(1–2), 131–141.

Kroll, Y., and L. Davidovitz (2003): "Inequality aversion versus risk aversion," *Economica*, 70(277), 19–29.

Kruger, J. (1999): "Lake Wobegon be gone! The" below-average effect" and the egocentric nature of comparative ability judgments.," *Journal of personality and social psychology*, 77(2), 221.

Larrick, R., K. Burson, and J. Soll (2007): "Social comparison and confidence: When thinking you're better than average predicts overconfidence (and when it does not)," *Organizational Behavior and Human Decision Processes*, 102(1), 76–94.

Malmendier, U., and G. Tate (2005): "CEO overconfidence and corporate investment," *The Journal of Finance*, 60(6), 2661–2700.

Moore, D., and P. Healy (2008): "The trouble with overconfidence," *Psychological review*, 115(2), 502–517.

Odean, T. (1998): "Volume, volatility, price, and profit when all traders are above average," *The Journal of Finance*, 53(6), 1887–1934.

Park, Y., and L. Santos-Pinto (2010): "Overconfidence in tournaments: evidence from the field," *Theory and Decision*, 69(1), 1–24.

Shepperd, J., J. Ouellette, and J. Fernandez (1996): "Abandoning unrealistic optimism: Performance estimates and the temporal proximity of self-relevant feedback.," *Journal of Personality and Social Psychology*, 70(4), 844–855.

Svenson, O. (1981): "Are we all less risky and more skillful than our fellow drivers?* 1," *Acta Psychologica*, 47(2), 143–148.

Van Dijk, W., M. Zeelenberg, and J. Van Der Pligt (2003): "Blessed are those who expect nothing: Lowering expectations as a way of avoiding disappointment," *Journal of Economic Psychology*, 24(4), 505–516.

Yates, J., J. Lee, and J. Bush (1997): "General Knowledge Overconfidence: Cross-National Variations, Response Style, and," *Organizational Behavior and Human Decision Processes*, 70(2), 87–94.