CONFIDENCE SNOWBALLING AND RELATIVE PERFORMANCE FEEDBACK

Zahra Murad
Economics and Finance, University of Portsmouth, UK, PO1 2UP
UNEC Cognitive Economics Centre, Azerbaijan State University of Economics, Azerbaijan

Chris Starmer*
School of Economics, University of Nottingham, UK, NG7 2QL

22/07/2021

Abstract

We investigate whether uninformative relative performance feedback can create biases in confidence leading it to 'snowball'. We study elicited confidence about own performance, relative to other group members, in three stages. As subjects move across stages, we change group composition so that new groups contain either only top performers or only bottom performers, from the previous stage. Between treatments, we manipulate whether subjects know about their own past relative performance or that of currently matched group members. In a treatment where subjects receive no feedback between stages, their confidence remains calibrated and stable across the stages. When subjects receive feedback in the other two treatments, their confidence snowballs in the direction of the feedback, both when feedback is fully informative and completely uninformative of their future performance. The results suggest the possibility of confidence biases emerging and snowballing in a potentially wide range of field settings.

Keywords: overconfidence, relative performance feedback, confidence updating

JEL: C91, D81

Acknowledgements: We thank Alexander Coutts, Robin Cubitt, Lingbo Huang, Boon Han Koh, Levent Neyse, Simone Quercia, Martin Sefton, Joep Sonnemans two anonymous referees and the editor for very helpful input to this paper. We also thank participants in seminars and workshops where this project has been presented for helpful comments and suggestions. This work was supported by the Economic and Social Research Council (grant numbers ES/K002201/1, ES/P008976/1). Declarations of interest: none.

* Corresponding author: chris.starmer@nottingham.ac.uk, Sir Clive Granger Building, University Park, University of Nottingham, Nottingham NG7 2RD

1 Introduction

Overconfidence is one of the most documented behavioral biases, and evidence for overconfidence has been found among students, athletes, entrepreneurs, and CEOs (Billett & Qian, 2008; Buser, Niederle, & Oosterbeek, 2014; Camerer & Lovallo, 1999; Daniel & Hirshleifer, 2015; Hilary & Menzly, 2006; Reuben, Sapienza, & Zingales, 2012; Rosenqvist & Skans, 2015). Yet the persistence of such overconfidence is puzzling, given that many individuals receive feedback on their abilities. While some studies have shown that feedback does reduce overconfidence (Moore & Cain, 2007; Murad, 2016; Ryvkin, Krajč, & Ortmann, 2012; Wozniak, Harbaugh, & Mayr, 2014), a wealth of empirical evidence suggests that individuals may be biased in processing feedback about their performance (Ambuehl & Li, 2018; Ertac, 2011; Möbius, Niederle, Niehaus, & Rosenblat, 2014). Critically, this evidence has focused exclusively on informative but sometimes noisy feedback. In this paper, we present a novel experiment which studies the evolution of confidence biases in the face of completely uninformative feedback.

Studying confidence as a response to uninformative performance feedback can be difficult with naturally occurring field data.[1] Confidence beliefs usually cannot be directly observed and may be confounded by factors such as self-selection; the interaction of confidence with effort and risk attitudes; and the effects of previous unobserved experiences. We thus use a controlled lab experiment to isolate the effects of uninformative relative performance feedback on confidence uncontaminated by other possible effects that are present in many naturally occurring settings. Moore and Healy (2008) distinguish between three types of confidence: estimation, placement and precision. We exclusively focus on placement where subjects probabilistically place themselves above or below the median. Given that individual-level confidence biases are prone to various statistical errors (Benoît & Dubra, 2011; Ryvkin et al., 2012), we focus on testing for systematic over- and underconfidence biases at group-level.

In the experiment, individual performances are compared within groups of four, which are re-matched over three stages. Top (bottom) performers are always re-matched with other top (bottom) performers. Given these changes in group composition, in our environment a rational agent who knows the matching rules should not alter their confidence levels, i.e. they

[1] Exception to this rule are the two studies by Murphy and Weinhardt (2020) and Denning, Murphy, & Weinhardt, (2020) which use exogenous variations of class matchings and ordinal rank that students receive on students' long term outcomes. They find significant effect of ordinal rank on students' confidence, performance, college participation, income and other variables.

should not become more (or less) confident as they are matched with other subjects similar to them. In our main treatment, subjects receive full feedback about the matching structure making the actual feedback completely uninformative about their future performance. However, we find that confidence does change systematically. Specifically, individuals who receive feedback that they were 'Top half performers' become more and more overconfident while the reverse is true for those told that they were 'Bottom half performers'. We call these observed behavioral responses to uninformative feedback 'confidence snowballing'.[2]

To benchmark this effect, we conduct two other treatments where feedback is either absent or fully informative about a subject's future performance. We find that when feedback is absent, subjects' confidence remains calibrated and stable across the stages. When feedback is informative, we find that confidence snowballs in the direction of feedback and more importantly we find very limited evidence that the degree of snowballing is different from when feedback is uninformative. Altogether, our results provide novel evidence that confidence biases can emerge even when feedback is completely uninformative. Our results are consistent with the reference group neglect hypothesis according to which subjects focus on feedback about their own performance and neglect information they receive about the relevant reference group.

Because the focus of our paper was to test, we believe for the first time, whether uninformative feedback can *cause* confidence biases, we aimed to create an environment where (at least at group level) the initial confidence assessments of own performance would be well-calibrated (i.e. no systematic confidence biases exist to start with). For this purpose, we choose the 'Circle task' introduced by Hollard, Massoni, & Vergnaud (2016). We discuss the task and its properties in more detail later (see Section 2 and Appendix A1). The crucial thing to note is that, as we conjectured and as evidenced by the data we report below, individuals' initial relative confidence assessments (prior to any performance feedback) were, on average, well-calibrated. To help convey the uninformativeness of the feedback to subjects, we set the difficulty of the Circle task to be either very easy or very difficult. This way, the possibility of being in the top or bottom half of one's group could be quite readily perceived as essentially random. The extant literature on confidence updating as a response to feedback has focused on environments that subjects already have some pre-existing feedback/experience in and tracks how feedback provided in the lab affects these a priori biased beliefs (such as overconfidence

[2] We use the label *snowballing* as a convenient metaphor to capture the idea of something becoming bigger (in absolute terms) as a consequence of some process evolving through time. In our case, we refer to biases getting bigger, which is applicable to both overconfidence and underconfidence bias.

or better/worse-than-average). These include beliefs about one's numeracy skills, beauty, IQ, general knowledge or driving skills among others (Berlin & Dargnies, 2016; Eil & Rao, 2011; Ertac, 2011; Möbius et al., 2014; Moore, 2007; Svenson, 1981). Contributing to this literature, we show that confidence biases can emerge and persist as a result of positively or negatively signed feedback even when the feedback is completely truthful but uninformative in nature. Choosing a task that was novel to subjects and not so ego-relevant as previously used tasks in the literature, we also provide a 'colder' environment for confidence snowballing uncontaminated by ego concerns (Drobner & Goerg, 2021; Eil & Rao, 2011; Ertac, 2011).

We contribute to the literature that studies how feedback affects confidence biases. Most of this literature has studied feedback as a tool to eliminate confidence biases. Moore & Cain, (2007), Murad, (2016) and Rose & Windschitl, (2008), Ryvkin et al. (2012) find that informative deterministic feedback is partially successful in eliminating confidence biases.[3] When feedback is informative but noisy, Coutts (2019), Ertac (2011) and Mobius et al. (2014) show that people treat positive and negative feedback asymmetrically and update beliefs insufficiently to de-bias their judgements completely. Burks et al. (2013), Eil and Rao (2011) and Ambuehl & Li, (2018) show that people exhibit dislike, or are indifferent to, new information when expecting negative feedback and hence they don't learn from it. Eberlein, Ludwig, & Nafziger, (2011) and Buser, Gerhards, & van der Weele, (2018) demonstrate heterogeneity in how conservative (or otherwise) individuals are in updating their confidence beliefs. We study the updating of relative confidence beliefs with respect to others in one's group, in a task where confidence is initially well-calibrated. In contrast to almost all existing studies examining how confidence biases are affected by *informative* relative performance feedback, we investigate whether *uninformative* relative performance feedback can induce confidence biases. So, while the rest of the literature shows that people update conservatively when they should be updating more, we provide evidence of a new updating anomaly: a case where people persistently update when they shouldn't be updating at all. The only other studies that look at the effect of uninformative feedback on confidence are by Murad and Huang (2020) and Banerjee, Datta Gupta, & Villeval, (2020). Both study an alternative form of uninformative feedback effect whereby feedback given in one task spills over to affect confidence beliefs and decisions in another unrelated task.

[3] More experiments on absolute confidence biases (overestimation) have also studied feedback as a remedy to cure overconfidence (Arkes, Christensen, Lai, & Blumer, 1987; Petrusic & Baranski, 1997; Pulford & Colman, 1997; Subbotin, 1996).

As far as we know, we are one of the first to use re-matching of individuals into groups to manipulate the informativeness of feedback. The only other paper that changes the reference group individuals compare themselves against over two rounds of feedback is by Berlin and Dargnies (2016). They mainly focus on gender differences in willingness to compete as a reaction to changing reference groups. We complement their findings that subjects update their beliefs when they shouldn't. However, they find that men and women update their confidence differently, possibly due to the strength of the gender stereotype of the task that they use (mathematical task). Differently from them, we use gender-neutral task uncontaminated by previous experience and perceptions. Hence, we can track confidence formation and evolution from the start in a neutral domain and provide a colder test of feedback effects on confidence.

2 Experimental Setup

Our experiment features three main treatments designed to test whether we can generate confidence snowballing as a result of uninformative feedback and, if we can, we wish to assess how far any snowballing observed is attributable to reference group neglect. In our environment, individuals complete a task in each of three stages and, in some treatments, we introduce feedback on subjects' own behaviour in the expectation that this may stimulate snowballing. As a benchmark, we include a treatment in which subjects receive no performance feedback at all: in this case, we expect no snowballing. We include two variants of feedback treatments. In both of these treatments, we provide subjects with feedback on their own performance by telling them (truthfully) that their own performance placed them, respectively, in either the top or the bottom half of their group in the previous stage. Between stages, however, we manipulate group composition such that if subjects knew this and fully took account of the group composition, they would realise that the feedback should not affect their own confidence. In one of the feedback treatments, subjects are fully informed about the changing group composition; hence, observing snowballing in this treatment is evidence of some degree of reference group neglect. To allow us to assess the extent of reference group neglect, we include a second feedback treatment designed to *induce* full reference group neglect – we do this by not providing information to subjects about the group composition. Conditional on finding snowballing, we test the extent of reference group neglect by comparing the degree of snowballing across the two feedback conditions. Full reference group neglect implies the same degree of snowballing in both treatments.

2.1 Feedback Manipulation

The experimental design consisted of three stages (within-subject) and three treatments (between-subjects). At each stage, subjects were matched into groups of four and had to complete a Circle Task (see the next subsection). In all treatments, whether subjects were told this or not, the group matchings were always adjusted between each stage such that any pair of subjects who had performed in the top (respectively bottom) half of their group went on to be grouped with another pair of subjects in the next stage who had also been in the top (respectively bottom) half of their group. Between-treatments we manipulated whether subjects received feedback about their own relative performance feedback and whether they knew about the composition of their group members, based on relative performance.

In the NoFeedback treatment, subjects did not receive any feedback (about their own performance or the changing group composition): after they completed a stage, they were just told that, in the next stage, two members (including them) from their old group would be matched with two members from another group. In the OwnFeedback treatment, after completing a stage they received either positive or negative feedback about their own relative performance: the feedback, which was true, simply read as "You were in the TOP two of your group" or "You were in the BOTTOM two of your group". They were then told that in the next stage, two members (including them) of their old group would be matched with two members from another group (without specifying the relative performance levels of their new group members).[4] In the FullFeedback treatment, after completing a stage, subjects received either positive or negative feedback of whether they were in the Top/Bottom two of their groups in the stage (similar to the OwnFeedback treatment). They were then told that the two Top/Bottom half scorers of their group (including them depending on their ranking) would be matched with the two Top/Bottom half scorers of another group in the next stage, respectively. Thus in the FullFeedback treatment, subjects knew whether their last round's performance placed them in the top or bottom half of the group and that they would be matched to similarly ranked subjects in the next stage. For example, if they received information that they were now in TopTop

[4] This design choice increases the informativeness of the feedback in the OwnFeedback treatment and thus predicts strong confidence snowballing. To see this, note that a subject receives feedback about their own relative performance in a group and knows that they are matched with one subject from their old group and two new subjects from another group. While they have no prior information about the new subjects from the other group, the subject that comes from their old group has a 1/3 chance of being of the same ranking and 2/3 of being of the opposite ranking. The converse is true for the FullFeedback treatment, subjects have symmetric information in both stages if they are grouped according to their performance with the members of another group. On average, subjects should have 50% confidence in Stage 1 and the positive feedback (Top half) about one's own performance should be cancelled out by the knowledge that they are matched with the Top half performers of another group. Note that this should be true assuming any performance score distributions, since the elicited confidence is about the median and not the mean.

group, they were reminded that all other three members of their new group had also been ranked as Top twice previously. Or, as a further example, if they received information that they were now in BottomTop group, they were reminded that all other three members of their new group had also previously ranked as Bottom after Stage 1 and Top after Stage 2.

The group matchings were explained to subjects in the instructions with the aid of diagrams to make the matching procedure as obvious as possible. Subjects' understanding of the matching structure at each stage of the experiment was a crucial element of our design. Hence, we reminded them both verbally and graphically about how the grouping was formed and what group they were in at the beginning of each stage to make the changing reference groups a salient feature of the experiment.[5] In the FullFeedback treatment, we included the words "Top" or "Bottom" in the group names to enhance salience. In the NoFeedback and OwnFeedback treatments, the Top/Bottom labels in the group names were simply replaced with numbers ("1" or "2"). For example, the top two scorers of group A would be matched with the top two scorers of group B, and the new group would be called group AB_Top in the FullFeedback treatment and group AB_1 in the NoFeedback and OwnFeedback treatments. After the second stage, the top two scorers of group AB_Top (AB_1) would be matched with the top two scorers of Group CD_Top (CD_1) to form the group ABCD_TopTop (ABCD_11). Similar matching was done for the bottom two scorers. Appendix B Figure B1 provides a summary of the matching structure.

2.2 The Task

The Circle Task consisted of a set of 20 circle pairs presented to subjects sequentially. In a single task, an individual subject saw a pair of black circles, on a computer screen, for 1 second. Each circle contained a number of white dots and the subject's task was to judge which of the two circles had the most dots (Hollard et al., 2016). For our purposes, the Circle task has several attractive features that we confirmed in a pre-test experiment consisting of 30 subjects (See Appendix A1 for details of the pre-test). We found that subjects were not generally familiar with the task. We also established that a subjects' guess of their score only minimally correlated with their actual score on the absolute level; and, importantly for our research objectives, there were no systematic biases in initial, relative confidence beliefs. This ensured that we did not have to worry about possible interaction effects between prior confidence biases and feedback effects. The pre-test data also showed no evidence of learning across stages,

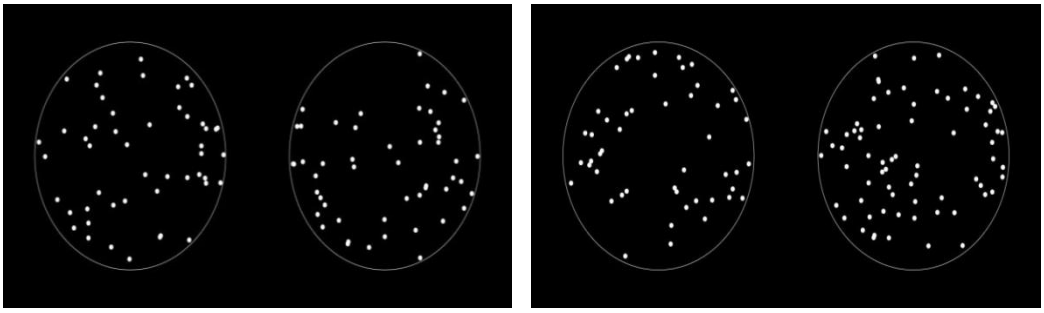
[5] This was also aimed at mitigating experimenter demand effects discussed at the end of Section 3.3.

limiting concerns that performance differences between stages might interact with confidence levels. By testing the impact of large variations in piece-rate payments for success in the task, we also established that performance was highly insensitive to incentives. This is important for our purposes in reducing the likelihood that snowballs in confidence could be attributed to perceived changes in incentives, arising across stages, for example from variation in own performance feedback. We also found that performance and confidence in relation to the task were unrelated to gender, risk attitudes and competitive attitudes.

Some previous studies have found the prevalence of overconfidence in easy and underconfidence in difficult tasks both in static and dynamic contexts (Hoelzl & Rustichini, 2005; Moore & Cain, 2007; Murad, 2016). With this in mind, in the main experiment, we manipulated task difficulty between-subjects: the Difficult variant had circles with 50&51 dots and the Easy variant had circles with 50&70 dots (see Figure 1). In the pre-test, the easy task was perceived as very easy (median rating 1.17 out of 7) and the difficult task was perceived as very difficult (median rating of 5.83 out of 7). Both variants of the task aimed at creating a minimal difference in perceived and actual performance to further strengthen the uninformativeness of feedback. The difficult variant having only 1 dot difference between circles was designed to create a perception that getting the task right is very difficult so performance is random (thus ending up in the top and bottom is random). The easy variant having 20 dot difference between circle aimed to create a perception that getting the task right was very easy so performance is at a ceiling for all subjects (thus ending up in the top or bottom is random). Confirming this, the average score was 11 (*s.d. = 2.01*) in the difficult and 19.79 (*s.d. = 0.49*) in the easy variant in the pre-test and 10.7 (*s.d. = 1.46*) and 19.7 (*s.d. = 0.82*) respectively in the main experiment (Figure C1 in Appendix C).[6]

At the start of the experiment, subjects read instructions then practised one set of 20 Circle tasks (within the same difficulty level as their incentivised tasks). The purpose of the practice set was to familiarise subjects with the task before eliciting Stage 1 confidence. No performance feedback was given to subjects for the practice set to avoid influencing confidence levels in Stage 1. After the practice set, subjects were informed about the matching structure of groups that would apply to them which varied by treatment, as explained above.

[6] Unexpectedly, the difficult variant of the task was more than random percentage of correct answers (significantly above 50% correct), meaning that there were some subjects able to even detect 1 dot difference between the circles accurately.



50&51 dots: Difficult 50&70 dots: Easy

[*After 1 second, the circles disappear from the screens*]

Which circle had more dots?
Left Right

Figure 1: The Circle Task

2.3 Confidence Elicitation

Each experimental session involved 16 subjects. In Stage 1, the 16 subjects were randomly matched into four groups of four people. Each group was given an identifier (Group A, B, C or D) and subjects were told which group they were in and that they all would be completing a set of Circle tasks similar to the practice set. Before starting Stage 1, they were asked to complete a table as in Figure 2. This is a tool designed to infer confidence from a set of choices.[7] For every row of the table, subjects had to make a choice between two options, Option A or Option B. Subjects knew that any choice from any table could be paid for real at the end of the experiment. As such, the interpretation of options is that Option A (if for real) would pay £10 if a subject was one of the top two of her group, according to her performance in the set of Circle tasks of that stage and £3 otherwise. Option B was a lottery which paid either £10 or £3 with the chance of £10 progressively decreasing down the rows. Given the construction of the table, subjects were expected to choose Option B in the first row and Option A in the last row. At some point, they were likely to switch from option B to A, and this switch point was used to measure each individual's confidence of being in the top half of their group

The elicited confidence in our design is interpreted as a subjective probability that a subject assigns to being in the top half of their group for a given stage. For example, subjects who assigned an equal probability to being in the top or bottom half of their group should

[7] Relative confidence is usually elicited via self-reports of relative percentile rankings (Brookins, Lucas, & Ryvkin, 2014; Buser et al., 2018; Coutts, 2019; Larrick, Burson, & Soll, 2007; Möbius et al., 2014; Moore & Healy, 2008), from estimates of absolute performance for self and others (Kamas & Preston 2012, Neyse, Bosworth, Ring & Schmidt 2015) or inferred from observed competitive payment schemes (Hoelzl & Rustichini 2005; Niederle & Vesterlund 2007). We choose to infer confidence from choices as it provides a more behavioural measure of confidence and is less prone to experimenter demand effects discussed at the end of Section 3.3.

switch in the 11th row of the table and would be imputed a 50% confidence (± 2.5%).[8] If we assume additive subjective probability distributions, then this task elicits the whole belief distribution of any subject. The tool also has the attraction of being incentive-compatible for confidence revelation under both expected utility theory and prospect theoretic models which allow non-linear probability distortion (see Urbig, Stauf, & Weitzel, 2009). The tool is robust to the Benoît & Dubra, (2011) critique of "apparent overconfidence" and meets the requirements for incentivising confidence elicitation proposed by Benoit, Dubra & Moore (2015) (see Appendix A2 for further details).[9]

After completing the confidence elicitation table at the beginning of a stage, subjects then undertook a set of 20 Circle tasks that determined their performance for the stage. Each correct answer was worth £0.50 hence subjects could earn up to £10 for the set.

Row	Option A: Bet	Your Choice	Option B: Lottery
1	You get £10 if you are one of the top two scorers of your group and £3 if you are one of the bottom two scorers of your group	A o o B	£10 with 100% chance
2		A o o B	£10 with 95% chance and £3 with 5% chance
3		A o o B	£10 with 90% chance and £3 with 10% chance
4		A o o B	£10 with 85% chance and £3 with 15% chance
5		A o o B	£10 with 80% chance and £3 with 20% chance
6		A o o B	£10 with 75% chance and £3 with 25% chance
7		A o o B	£10 with 70% chance and £3 with 30% chance
8		A o o B	£10 with 65% chance and £3 with 35% chance
9		A o o B	£10 with 60% chance and £3 with 40% chance
10		A o o B	£10 with 55% chance and £3 with 45% chance
11		A o o B	£10 with 50% chance and £3 with 50% chance
12		A o o B	£10 with 45% chance and £3 with 55% chance
13		A o o B	£10 with 40% chance and £3 with 60% chance
14		A o o B	£10 with 35% chance and £3 with 65% chance
15		A o o B	£10 with 30% chance and £3 with 70% chance
16		A o o B	£10 with 25% chance and £3 with 75% chance
17		A o o B	£10 with 20% chance and £3 with 80% chance
18		A o o B	£10 with 15% chance and £3 with 85% chance
19		A o o B	£10 with 10% chance and £3 with 90% chance
20		A o o B	£10 with 5% chance and £3 with 95% chance

Figure 2: Confidence Elicitation Table

[8] We eliminated the possibility of double switching by presenting subjects who attempted to double switch with an error message that asked them to think more carefully about their choices and decide on their switching point.
[9] Our elicitation is also robust to the issues raised by Krajc and Ortmann (2008) to explain statistical reasons why subjects can be found to be over/underconfident. See Appendix A2.

2.4 Procedures

After Stage 3, subjects completed a questionnaire self-reporting socio-demographic information, risk attitudes, competitive attitudes and their general level of confidence (Dohmen, Falk, Huffman, & Sunde, 2010). In the questionnaire, we also asked subjects whether their eyesight was impaired. Once all subjects had completed the questionnaire, the experimenter approached each subject with randomization devices to, privately, determine their payment. For each subject, one of the stages was randomly selected and within the chosen stage either one of the rows of the confidence elicitation table or the piece rate performance in a set of Circle tasks was selected. We used physical randomization devices (dice and a bag of numbered balls) and explained this in the instructions to help make the independence of the randomization as clear as possible to subjects. Full instructions of the experiment are in Appendix B with a summary of the experimental flow in Appendix C, Table C1.

Subjects were 384 undergraduate and postgraduate students recruited via Orsee (Greiner, 2015) at the CeDEx laboratory, University of Nottingham. Subjects, 54% female with 18% from Economics and Business majors, were divided into three treatments with 12 sessions for the FullFeedback treatment and 6 sessions each for the OwnFeedback and NoFeedback treatments.[10] The experiment was programmed using Ztree (Fischbacher, 2007) and lasted around 30 minutes. Subjects were seated in private cubicles and were given paper instructions. The instructions were read by the same experimenter at the beginning of the experiment. The average subject payment was £7.60 (min. £3; Max £10).

2.5 Hypotheses

Here we set out our main expectations regarding snowballing effects within and between treatments. Our first hypothesis is that there should be no *systematic* change of confidence in the FullFeedback treatment. This is not to deny that it could be rational for an *individual* to change their confidence upwards or downwards after receiving feedback about their new reference group and their past performance. At an individual level, subjects may have incorrect (or imprecise) initial confidence assessments that they report in the first stage and

[10] Given that the feedback environment is richest in the FullFeedback treatment and there are a number of opposing feedback effects that can affect confidence (information about own performance and information about the group members), we assumed small effect sizes and calculated the required number of observations powered at 80% with 5% error probability. We then adjusted the required number of observations in the NoFeedback treatment where we expect no effect (where the lower or higher sample size does not influence statistical power), and medium effect sizes in the OwnFeedback treatments (where the lower sample size provides similar statistical power). This resulted in required minimum of 48 subjects per feedback condition (i.e. TopTop/BottomBottom) in the FullFeedback treatment and a minimum of 24 subjects per feedback condition in the OwnFeedback treatment.

adjust in later stages as they experience the task. This is especially the case in our setting given that subjects have little to no previous experience of the task and hence are unaware of their skills. However, at *group level* there is no reason why rational updating of individual confidence would lead to a systematic pattern of snowballing in one direction or the other as subjects are matched with similar others across the stages. Our first hypotheses is thus that *no confidence snowballing should occur in the FullFeedback treatment*.

Now consider differences in the informational structure between the treatments. In the OwnFeedback treatment, since subjects receive feedback only about their own performance, a rational individual should update their confidence in the direction of the feedback received: that is, they should decrease their confidence after learning that they are in the Bottom half and increase their confidence after learning that they are in the Top half of performers. Since all rational subjects should do this, *we expect to observe systematic snowballing in the OwnFeedback treatment* in the direction of feedback received. Conversely, *there should be no systematic snowballing in the NoFeedback treatment* since there is no information provided to the subjects that they can use to, rationally, update their confidence. Based on these hypotheses regarding rational confidence updating, we may also test two further hypotheses: *there should be significant difference in snowballing patterns between the FullFeedback and OwnFeedback treatments*; while *we should observe no systematic difference between FullFeedback and NoFeedback treatments*.

3 Results

We first present descriptive results on the distribution of prior confidence levels. We then present evidence of confidence snowballing and the emergence of confidence biases from one stage to the next for the subjects that received relative performance feedback. We check the robustness of confidence snowballing and test for reference group neglect by comparing the FullFeedback and OwnFeedback treatments. In addition, we run tests for the existence of other confidence updating biases in our data such as fundamental misattribution bias, good-news-bad-news effects and motivational feedback effects. Finally, we discuss the possibility of experimenter demand effect driving our results.

3.1. Stage 1 Confidence: is confidence well-calibrated prior to performance feedback?

We first investigate whether initial beliefs about task performance are well-calibrated, prior to any feedback on task performance. Regarding the rational benchmark for belief formation, whilst individuals may estimate their ability with error (e.g. overconfidence or

underconfidence), the distribution of these errors should have a mean of zero. Following Benoît & Dubra, (2011), in this benchmark case we require that beliefs of scoring in the top two of a group of four (i.e. being in the top 50%) are on average equal to 50% (± 2.5%). As Benoît & Dubra, (2011) show in their Theorem 3, if this does not hold in the population, then such beliefs cannot be rationalized.

Over all treatments, the median and modal first stage confidence in our data is 50%, and the mean is 52.29% (*s.d. = 17.02, n=384)* suggestive of initially well-calibrated beliefs. This impression is confirmed by the Wilcoxon sign-rank test, which fails to reject the null hypothesis that median confidence is 50% in the NoFeedback and in the FullFeedback treatments (median = 52.5% and 50.0% respectively, *p-value* > 0.10). While there is some evidence of overconfidence in the OwnFedback treatment (mean = 55%, *p-value* = 0.096), it is relatively weak compared to previously cited studies (e.g. 63% and 56% in Hoelzl & Rustichini, (2005) and 67% in Benoît, Dubra, & Moore, (2015). The cumulative density functions of elicited confidence in Stage 1 are presented in Figure 3, separately for each treatment.

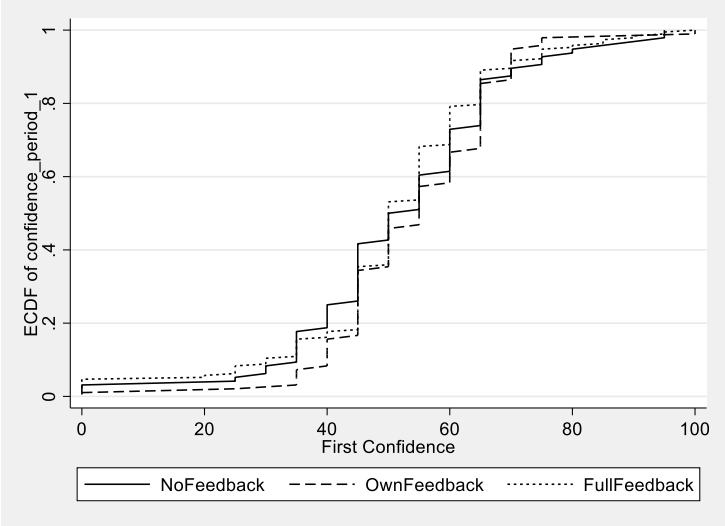


Figure 3: CDF of Stage 1 Confidence across the treatments

Eyeballing of Figure 3 suggests no obvious cross treatment differences in the distributions of stage one confidence. Table 1 provides further tests for cross-treatment differences in first-stage confidence via regression analysis that allows us to exploit additional controls including task difficulty and individual-level characteristics elicited from the survey measures. Consistent with the extant literature on relative confidence biases, we find that the first stage confidence is lower in the difficult variant of the task than in the easy variant of the

task when we look at all the treatments combined.[11] The magnitude of the difference is quite small compared to other studies reporting over/underconfidence in easy and difficult tasks. We conjecture that it is due to the task being one where subjects do not have a-priori systematically miscalibrated perceptions or beliefs about their own or others' performances (see more discussion of this in Appendices A2 and A3). In regression analysis, we find that none of the individual-level characteristics, including gender, are significant predictors of confidence.[12] In terms of treatment differences, we find no evidence that first stage confidence in the NoFeedback treatment is different from confidence in either of the feedback (FullFeedback or OwnFeedback) treatments. We do, however, find a marginally significant difference between the OwnFeedback and FullFeedback treatments; post-estimation Wald statistics is equal to -3.65 (*p-value* = 0.060) without controls and -3.39 (*p-value* = 0.089) with controls. Given these results, in the parametric analysis of feedback effects that follow, we always control for Stage 1 confidence levels and the difficulty level of the task. We also find no correlation between first stage confidence levels and receiving Top feedback: hence the starting point for Top and Bottom participants is identical in terms of prior confidence levels.

3.2. Confidence Snowballing and the role of reference group neglect

Our main research question is how confidence changes from one stage to the next in response to the relative performance feedback (and changing reference groups). The main results are summarised in Figure 4. The figure presents the mean confidence of subjects at each stage with four separate panels for different pairings across rounds (TopTop, BottomBottom, TopBottom and BottomTop). Since we find no significant differences in snowballing behaviour across difficulty levels (see Appendix A3: Table A4), in the snowballing analysis we pool data across the two difficulty levels.[13] Within each panel, we report pairwise tests for equality of confidence across stages taking subject-level analysis, organized by treatment and which group the subject ended up in. We use Wilcoxon matched-pairs sign-rank tests that do

[11] When we test for significance of difference between difficulty levels in each treatment separately in Appendix A3, we do not find significant difference in confidence levels between the easy and difficult tasks possibly due to the number of observations being lower. Table A3 in Appendix A tests for interaction effects between the feedback treatment variables and difficulty levels: we find no differential effect of difficulty of the task on confidence across the three treatment conditions.

[12] The absence of gender differences in confidence is consistent with previous literature that has shown that the perceptions about the task being male or female can affect confidence levels and competitive preferences (Cárdenas, Dreber, von Essen, & Ranehill, 2012; Große & Riener, 2010; Günther, Ekinci, Schwieren, & Strobel, 2010).

[13] We report level differences in snowballing behaviour in the Appendix A3 of the paper and discuss the reasons behind the possible effects of the difficulty on confidence relating it to previous studies.

not require independence assumptions to test for pairwise differences as each pair is matched within-subject and controls for subject-level differences.

We highlight several patterns evident in Figure 4. First, looking across the four panels for the NoFeedback condition, it is apparent that there are no significant changes in confidence levels across stages. Hence, in the absence of feedback, confidence levels are stable (subjects behave rationally).

Table 1: Predicting First Stage Confidence Levels

DV: Stage 1 Confidence	**Model 1**	**Model 2**
OwnFeedback	2.34 (2.33)	2.57 (2.36)
FullFeedback	-1.30 (2.24)	-0.82 (2.29)
Top	0.44 (1.73)	0.25 (1.75)
Difficulty		-3.74 (1.73)**
Female		0.18 (1.79)
Age		0.48 (0.58)
Confidence_general		0.31 (0.65)
Competitive_general		0.03 (0.63)
Risk_general		-0.44 (0.72)
Eyesight		0.01 (1.78)
Constant	52.12 (1.88)	44.08***
N	384	384
R-Sq	0.008	0.025
Wald coef. Own vs Full feedback	-3.56*	-3.39*

*Risk, Confidence and Competitive are self-reported measures of willingness to take risks, how confident and how competitive a person is in general. * p < 0.10, *** p < 0.01 of the underlying standard errors are clustered at session level. Top is a binary variable if a subject was ranked Top or Bottom at the end of Stage 1.*

We next consider whether there is evidence of snowballing. The top two panels are those where we can test for snowballing effects of feedback because these are the cases where the feedback had a consistent direction across stages: feedback on own relative performance was either indicative of consistently top performance (TopTop) or bottom performance (BottomBottom); likewise, in the FullFeedback treatment, the feedback on the reference group either indicated that the reference group was consistently populated with last rounds' better performers (TopTop) or consistently with the lower performers (BottomBottom). Looking at these top two panels, first, consider the OwnFeedback condition – and recall that these data are the cases where we have induced full reference group neglect (by not telling subjects how their

reference group was changing).[14] For these two cases, *we see clear and highly significant evidence of snowballing*: when feedback on own performance is consistently positive, average confidence consistently increases; but when feedback on own performance is consistently negative; confidence consistently falls.

Similar trends are apparent in the FullFeedback treatment, where subjects know that their reference group is changing, with systematic increases in confidence for TopTop and reductions in confidence for BottomBottom. This is evidence that subjects are prone to *some degree* of reference group neglect. The data reported in the lower two panels of Figure 4. – i.e. for BottomTop and TopBottom and relating to subjects who received two different signals between the two stages - tell a similar story: when subjects receive feedback on their own performance, confidence moves in the direction of the feedback, though because of the mixed nature of the feedback, we do not observe a snowball between the first and last stages. We thus reject our first hypothesis that there is no systematic snowballing in the FullFeedback treatment. Although highly significant, at the eyeball level, the snowballing effects in FullFeedback appear smaller in magnitude than those in OwnFeedback, consistent with partial rather than full reference group neglect.

To test whether reference group neglect is full or partial, we use linear regression to test for a difference in snowballing between the OwnFeedback and FullFeedback treatments. The results are presented in Table 2 where we report separate regressions for TopTop and BottomBottom. In this analysis, we control for a number of other subject-level characteristics and cluster errors at session-level. We test for the equality of confidence snowballing from Stage 1 to Stage 3 between FullFeedback and OwnFeedback treatments using two model specifications (Model 2 adds controls: individual characteristics; performance score; and task difficulty). For all four models we find that, as expected, confidence snowballs from Stage 1 to Stage 3 in the direction of the feedback received: it increases after receiving two consecutive positive feedbacks (statistically significant positive coefficient of Stage 3 in TopTop) and decreases after receiving two consecutive negative feedbacks (statistically significant negative coefficient of Stage 3 in BottomBottom). Testing for difference-in-difference in confidence snowballing using an interaction term, we find that the degree of snowballing is lower in the FullFeedback compared to the OwnFeedback treatment: the coefficient of Stage3×FullFeedback interaction term is negative in the TopTop and positive in

[14] This is not to say that subjects had no idea who their new group members were. In Footnote 5, we note that subjects may get some probabilistic information about one of their new group members by knowing their own ranking in previous group and that one of the old group members will be joining their new reference group.

BottomBottom conditions and shows around 6.5% points difference in confidence. However, neither coefficient is significant. This suggests that while there is some evidence for the second hypothesis in our data, it seems to be only partial as we observe significant differences in the pattern of snowballing between NoFeedback and FullFeedback treatments.

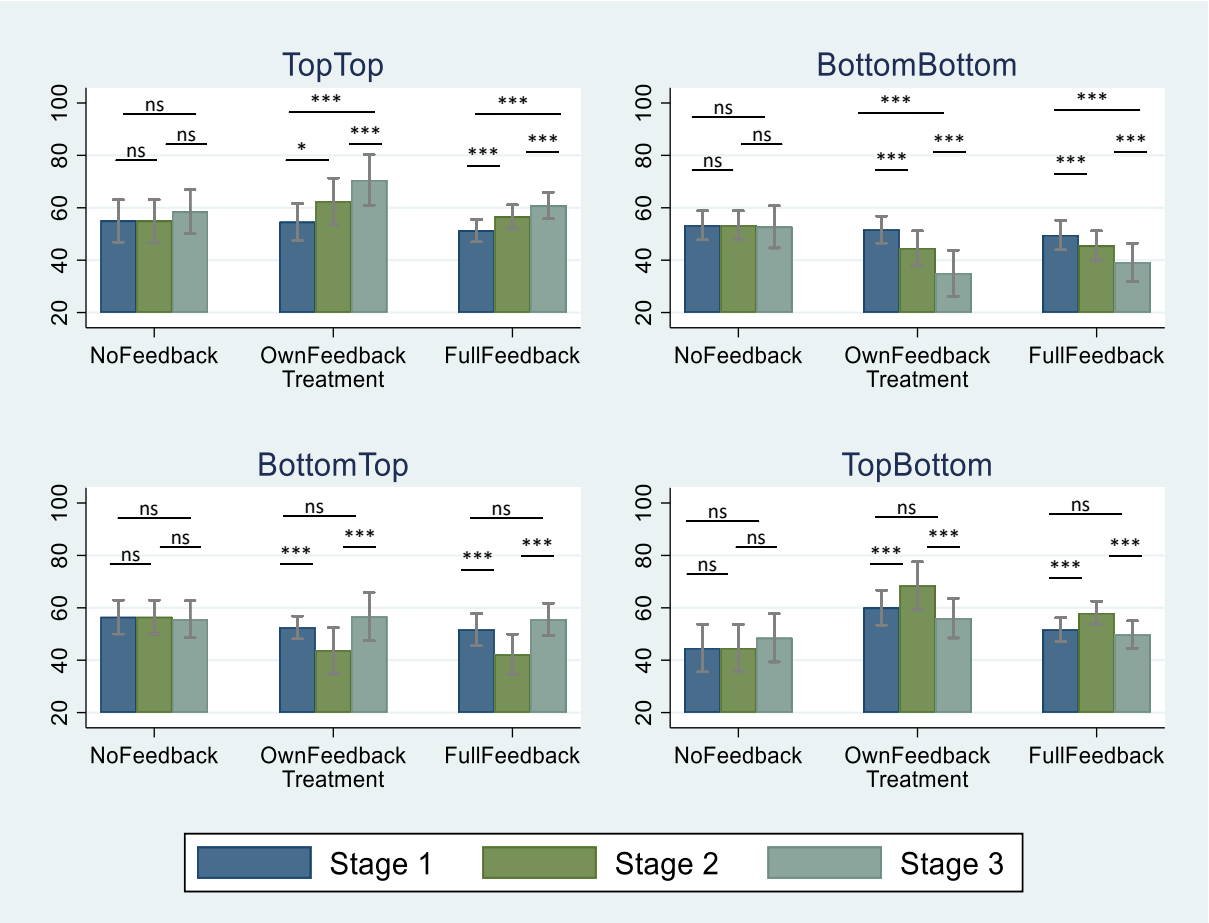


Fig 4. Mean individual confidence levels across stages and treatments. Wilcoxon matched-pair signed-rank test at an individual level. *ns p>0.10, * p < 0.10, *** p < 0.01. The error bars are ±SEM.*

To complement this analysis, we also pairwise test whether the absolute level of confidence is significantly different between the OwnFeedback and FullFeedback treatments, in each stage, using Wilcoxon ranksum tests. While the difference is significant in the case of the TopTop feedback condition, there are no differences in confidence levels for BottomBottom feedback condition between the treatments. For TopTop: Stage 1 p-value = 0.114; Stage 2 p-value = 0.073, Stage 3 p-value = 0.001. For BottomBottom: Stage 1 p-value = 0.880, Stage 2 p-value = 0.674, Stage 3 p-value = 0.345. Hence, in line with eyeball

impressions from Fig. 4, this provides some statistically significant evidence that reference group neglect is partial in some cases (i.e. the TopTop condition).

Table 2: Testing for complete versus partial reference group neglect

DV: Confidence	**TopTop**		**BottomBottom**	
	Model 1	Model 2	Model 1	Model 2
Stage 3	16.04***	15.98***	-16.66***	-18.51***
	(3.82)	(4.01)	(4.83)	(5.73)
FullFeedback	-3.33	-3.29	-2.08	-3.74
	(3.59)	(3.42)	(4.63)	(4.68)
Stage3×FullFeedback	-6.46	-6.34	6.25	7.05
	(4.38)	(4.53)	(7.05)	(7.68)
Constant	54.58***	36.96	51.66***	39.91
	(2.98)	(25.91)	(3.38)	(31.78)
Adj R^2	0.116	0.152	0.071	0.119
N	144	144	144	144
Controls	No	Yes	No	Yes

** 10%, *** 1% significance level, controls include actual performance score, task difficulty, gender and nationality dummies, age, general willingness to take risk, general level of confidence and competitiveness and whether any issues with eyesight was reported. Standard errors clustered at the session level.*

We conduct additional analysis of whether confidence snowballing following a positive (negative) feedback is stronger for participants who found it very unlikely to be top (bottom). Given the distribution of first stage confidence, we use 25th and 75th percentile points as thresholds for classifying subjects as "not confident" or "confident". These are subjects who had confidence levels of 45% or lower or 65% or higher, respectively. We find that in the FullFeedback treatment, subjects who were confident did not react to consecutive positive feedback as much as subjects who were not confident: snowballing was 10% points more for not confident subjects than for confident subjects (p-value = 0.082). For subjects who received two consecutive negative feedbacks, we do not find a significant difference in snowballing between confident and not confident subjects (p-value = 0.136). Details of this analysis are presented in Appendix C, Table C3 where we also estimate the triple interaction effect between feedback treatments, snowballing and initial confidence of subjects (see Model 2 of Table C3). With respect to the triple interaction effect, we do not find any significant differences between confidence snowballing across treatments and subjects being confident or not. This evidence, that there may be differences between how subjects' confidence snowballs depending on the initial level of confidence, is only suggestive, however, given our sample size and low

statistical power. Moreover, there may be statistical reasons why we observe higher confidence snowballing following positive feedback for not confident subjects compared to confident subjects: there is more room for confidence adjustment for not confident subjects than for confident subjects after positive feedback.

3.3 Is there evidence consistent with other confidence-updating hypotheses?

In this section, we test for alternative confidence updating mechanisms that have been previously studied in the literature. Firstly, feedback effects on confidence snowballing could be motivational (changing effort and thus performance) rather than simply cognitive (changing beliefs).[15] Feedback could affect subjects' subsequent motivations to exert effort and hence their confidence could follow this shift in their efforts/performance. We rule out this explanation for confidence snowballing. Firstly, our task was pre-tested so that incremental effort level exerted on the task did not improve/diminish performance (as long as the minimal effort was exerted; see Appendix A1). Secondly, in our analysis in Table 2, we explicitly control for the performance variable (the score in the set of Circle tasks in each stage). Hence any change observed in confidence from Stage 1 to Stage 3 is conditional on any change in performance. Thirdly, to test for other unobserved motivational effects of feedback on exerted effort levels, we asked subjects in the post-study questionnaire "Did you try harder in some stages of the Circle task than in others?" Subjects could choose "I tried equally hard throughout"; "I tried harder in Stage 1"; "I tried harder in Stage 2"; "I tried harder in Stage 3". 87%, 65% and 66% of our subjects in NoFeedback, OwnFeedback and FullFeedback treatments, respectively, reported that they tried equally hard throughout (Table 3). The responses of the remaining subjects were equally distributed across the stages. Moreover, no systematic relationship was detected across the feedback conditions (TopTop, TopBottom, BottomTop or BottomBottom) and subjects' answer to this question (χ^2 *p-value* > 0.650). We thus consider it unlikely that the motivational effects of feedback may have contributed to confidence snowballing. This assures us that our aim of isolating feedback effects on confidence are not contaminated by performance effects and uninformative feedback exogenously affects confidence.

[15] Ederer, (2010) and Fishbach, Eyal, & Finkelstein, (2010) argue for the motivating effects of positive feedback on performance and effort levels. Berger & Pope, (2011); Eriksson, Poulsen, & Villeval, (2009); Kuhnen & Tymula, (2012) show motivating effects of negative performance feedback on the subsequent effort levels in tournaments. Gill & Prowse, (2012); Malueg & Yates, (2010), argue for discouraging effects of negative performance feedback on the subsequent performance.

Table 3: Did you try in one stage harder than in the other? (frequencies of choosing an option)

FullFeedback	TopTop	TopBottom	BottomBottom	BottomTop	Total
I tried hard throughout	38	34	28	27	127
Stage 1	3	5	9	3	20
Stage 2	0	3	1	12	16
Stage 3	7	6	10	6	29
OwnFeedback					
I tried hard throughout	14	11	12	17	54
Stage 1	4	4	2	4	14
Stage 2	1	3	3	3	10
Stage 3	5	6	7	0	18
NoFeedback					
I tried hard throughout	-	-	-	-	83
Stage 1	-	-	-	-	3
Stage 2	-	-	-	-	6
Stage 3	-	-	-	-	4

Secondly, *causal misattribution bias* may have contributed to the observed confidence snowballing. Causal misattribution bias manifests itself through attributing successes to internal factors such as skill, personality and expertise and failures to external factors such as luck and other people's faults (Mezulis, Abramson, Hyde, & Hankin, 2004). At the end of the experiment, we asked subjects the question, "On a scale of 1 to 7, did luck or skill determine your performance in the Circle Task? (1 entirely luck, 7 entirely skill)". Figure 5 presents evidence of causal misattribution bias. For subjects who received feedback (left-hand panel), when in the TopTop condition they rate the task as more influenced by skill (*mean (sd) = 4.61 (2.01)*) compared to the ones in the BottomBottom condition (*mean (sd) = 3.04 (1.73)*); this difference is significant at *p-value* = 0.000 according to Wilcoxon ranksum test. In the additional analysis reported in Table 4, we control for the effect of these misattributed beliefs on confidence snowballing and check for any interaction effects for those who attributed feedback to skill versus luck. We find that there is marginally significantly higher snowballing from positive feedback for those who misattributed feedback more to skill than luck; this significance disappears, however, when we control for additional individual characteristics (socio-demographic and economic preferences). No similar result holds for those receiving negative feedback. So, while some causal misattribution bias exists in our data, we do not find that it is an important contributor to confidence snowballing in our study.

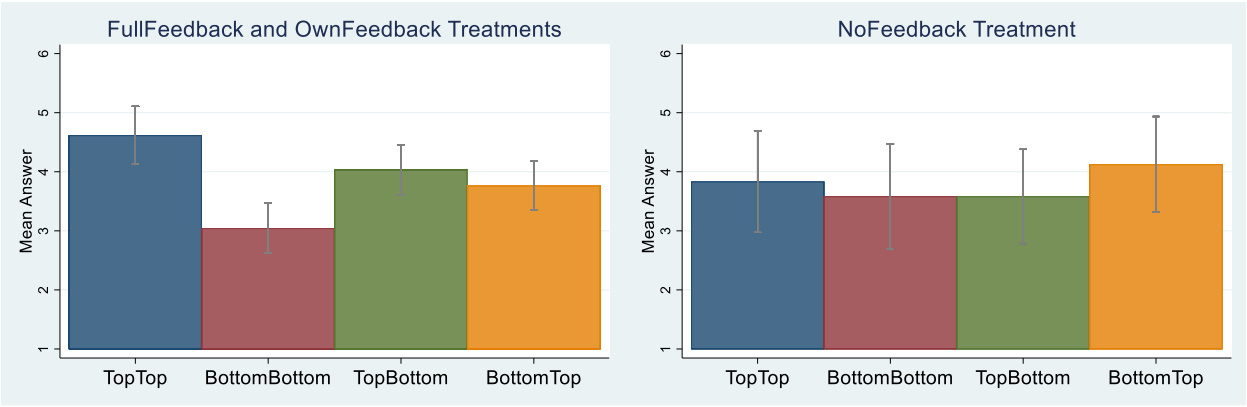


Figure 5: Causal Misattribution to Skill or to Luck depending on the feedback received
Mean answers to the question: On a scale of 1 to 7, did luck or skill determine your performance in the Circle Task? (1 entirely luck, 7 entirely skill)".

Table 4: Snowballing Confidence and Causal misattribution

The dependent variable is the change in confidence from the first to the last stage: Confidence Stage 3 - Confidence Stage 1

	Model 1	**Model 2**
Skill_Attr	-.208 (3.06)	-.44 (2.97)
TopTop	6.28 (2.89)**	6.71** (3.18)
BottomBottom	-14.75*** (4.72)	-16.52*** (4.42)
TopTop×SkillAttr	7.44* (3.72)	4.37 (3.91)
BottomBottom×SkillAttr	5.08 (7.61)	7.23 (7.02)
Constant	0.833 (2.70)	-34.00** (14.54)
Controls	No	Yes
N	288	288

** p <0.10, ** p < 0.05, *** p < 0.01; standard errors are in parenthesis and are clustered at session level. Includes controls such as gender, age, self-reported competitiveness, risk-seeking and self-reported confidence. Skill_Atrr is a dummy taking value 1 if above median. Pooled results for OwnFeedback and FullFeedback treatments.*

Next, we test for asymmetric updating behaviour depending on the feedback received. Positive asymmetric updating (also known as a *good-news-bad-news effect*) predicts overweighting of positive feedback and underweighting of negative feedback when updating beliefs (Eil & Rao, 2011; Möbius et al., 2014; Moutsiana, Charpentier, Garrett, Cohen, & Sharot, 2015). Positive asymmetric updating would predict that the change in confidence from Stage 1 to Stage 3 would be lower in magnitude after receiving negative feedback than after receiving positive feedback, with the opposite pattern for negative asymmetric updating. We

test for asymmetric updating by regressing the absolute change in confidence on the feedback condition dummy (TopTop vs BottomBottom) and results are reported in Table 5. Contrary to the good-news-bad-news effect, we find some evidence of negative asymmetric updating with a tendency for subjects to update less in the TopTop condition compared to the BottomBottom condition; see the negative coefficient on the treatment dummy in Table 5 which is significant at the 5% level in both specifications.

Our design also permits another form of test for differential responses to positive and negative feedback by examining the effects of mixed feedback in the TopBottom or BottomTop conditions. If asymmetric updating were in operation, we would observe a significant change in confidence in both of these conditions. As demonstrated in the bottom panels of Figure 4, there is no change in confidence from Stage 1 to Stage 3 in either the TopBottom or in BottomTop conditions of either of the feedback treatments. Hence in these conditions, we observe no asymmetry in updating.

Table 5: Testing for positively asymmetric updating

DV: Absolute value of (ConfStage3-ConfStage1)	Model 1	Model 2
TopTop (vs BottomBottom)	-6.18** (2.68)	-6.06** (2.75)
Constant	20.97*** (2.24)	-17.39 (14.80)
Controls	No	Yes
N	144	144

** p <0.10, ** p < 0.05, *** p < 0.01; standard errors are in parenthesis and are clustered at session level. Includes controls such as gender, age, self-reported competitiveness, risk-seeking and self-reported confidence. Pooled results for OwnFeedback and FullFeedback treatments.*

This adds to the mixed evidence relating to asymmetric information processing with recent studies showing either no asymmetry (Barron, 2020; Buser et al., 2018; Gotthard-Real, 2017; Grossman & Owens, 2012), positive asymmetry (Eil & Rao, 2011; Möbius et al., 2014; Moutsiana et al., 2015; Sharot, Korn, & Dolan, 2011) or negative asymmetry (Coutts, 2019; Ertac, 2011). There is more recent evidence that perceived ego-relevance of the task may affect the asymmetry in information processing. Drobner & Goerg, (2021) find more evidence of the good-news-bad-news asymmetry when the task is described as an IQ task than when not. We speculate that our task is much less ego relevant than those tasks usually used in the literature and this could explain why we do not detect higher confidence snowballing with positive, relative to negative, feedback. That said, we cannot rule out some role of ego-relevance given

our previous findings on casual misattribution bias: if subjects perceived the task as completely non-ego-relevant then there would have been no difference in the attribution of performance to skill or luck between the feedback conditions.

Finally, we consider whether some form of experimenter demand effect (EDE) (Zizzo, 2010) might be driving our results as a consequence of our design somehow encouraging or prompting subjects to adjust their behaviour, consistent with snowballing. If some form of EDE is an issue in our design, we think it quite likely that it would be a broader problem across studies testing belief updating behaviour given feedback (e.g. Buser et al., 2018; Coutts, 2019; Eil & Rao, 2011; Ertac, 2011; Grossman & Owens, 2012; Möbius et al., 2014) and quite plausible that such effects might be better controlled in our design, relative to some others. A key reason for thinking this is that, in our design, confidence is inferred from *incentivised choice behaviour* in a task involving quite high incentives (earning at least £3 and up to £10) for honest and careful confidence reporting and, hence, significant costs for mis-reporting. Moreover, subjects would arguably require a sophisticated analysis of the confidence elicitation task to form any view of how higher or lower switch points might be judged favourably by the experimenter. In the end, however, it is hard to rule out EDEs completely in almost any experimental context. Because of this, in our view, a persuasive argument that EDE's may be responsible for patterns in our data requires more than simply an assertion they could be. In our context, we see no obvious case for attributing the patterns to EDE's and some reasons, presented above, to be sceptical about that, especially in the context of our incentivised approach to confidence elicitation. Were some specific EDE hypothesis to be put forward as an account of our data in the future, however, that could and should be tested against our interpretation.

4 Concluding Discussion

We contribute to the literature studying confidence biases and how feedback affects them even when the feedback is uninformative in nature. Some other recent studies, like ours, also demonstrate impacts of uninformative feedback on confidence. Banerjee et al., (2020) examine how feedback spills over from one task to another unrelated one, affecting individuals' confidence and competitiveness. Huang & Murad (2020) used the same Circle task as us and showed that the feedback received in this task affects confidence and preferences for social comparison in a subsequent arithmetic task. Neither of these studies, however, identify the main result that we have presented here: that is, that *uninformative feedback on a task can create confidence biases that snowball*.

In many environments, the link between confidence and other factors such as motivation or ego relevance may be strong. How emergent confidence biases affect or interact with such other factors goes beyond the scope of this paper but there is evidence on this elsewhere (Drobner& Goerg 2021; Chen & Schildberg-Hörisch, 2019; Denning et al., 2020; Fischer & Sliwka, 2018; Fischer & Wagner, 2019; Murphy & Weinhardt, 2020). In our experiment, we aimed to isolate the effect of uninformative feedback on confidence and, to this end, we chose the Circle task in the expectation that the influence of motivation and ego-relevance would be minimal. While this was appropriate for our research objectives, it is important to recognise that such factors are at work more generally. For example, recent studies suggest that feedback may be processed differently depending on how ego-relevant a task is perceived to be (Castagnetti & Schmacker, 2020). As such, we see a useful space for future work to examine interactions between factors such as ego-relevance and the confidence updating process.

The snowballing mechanism we have identified is potentially relevant for behaviour in a wide range of naturalistic environments. For example, consider new students starting at college or financial traders starting a new job. To begin with, they may have little basis for judging their relative abilities (Gervais & Odean, 2001) but, over time, they receive relative performance feedback both about themselves and their reference groups: e.g. students being able to observe both their own and their peers' grades/performances; traders being exposed to their own and to market performance information. In such cases, our results point to the possibility of confidence biases emerging. For example, in work environments, employees and firms may underweight the feedback they receive about their competitors, responding mainly to feedback about their own performance, in potentially suboptimal ways (e.g. in relation to pricing, marketing strategies or promotions decisions). Similarly, students will become discouraged and underconfident when finding themselves in classes with extremely able peers, neglecting the fact that their reference group is more skilled than average. The opposite will be true for those students that are in classes with low ability peers (see the 'Big-Fish-Little-Pond effect' of Marsh, 1987). This may, in turn, affect their motivation and future performance (Denning, Murphy, & Weinhardt, 2020; Fischer & Sliwka, 2018; Murphy & Weinhardt, 2020).

We conjecture that the snowballing effect we have identified may also help to explain the *emergence* of some well-attested confidence anomalies such as systematic overconfidence or the better-than-average effect (Alicke & Govorun, 2005). For example, a frequently cited case of confidence miscalibration is the systematic overconfidence among drivers, most of

whom rate themselves as above average in ability (Svenson, 1981).[16] To see the potential connection with snowballing, picture the typical driver who most days completes their car journey to work or the school gates without incident. This regular, positive, accident-free, experience may be a source of growing confidence in their own relative ability, notwithstanding that most drivers, most days, receive essentially the same positive signal and become similarly more experienced as drivers. On this interpretation, this common bias may be a case of confidence snowballing (in an ego-relevant task), driven by uninformative feedback and reference group neglect.

On the assumption that they exist, how could tendencies towards confidence snowballing be mitigated? In many cases it may not be possible or desirable to completely 'turn off' the feedback that - we conjecture - is a driver of confidence snowballing. Nevertheless, there may be scope for those in relevant positions of responsibility (e.g., as managers, teachers, officials of government departments, etc.) to structure and manage feedback in ways that may moderate snowballing. For example, Zimmerman (2020) shows that recall of positive and negative information plays a significant role in how people update beliefs. Drawing on evidence of this sort, it may be possible to develop strategies for managing the recall of feedback, by target individuals, through its timing, frequency or framing. Individuals might also be explicitly coached in the art of learning how to place appropriate weight on information about their own performance versus information about the qualities of a relevant reference group. Our paper suggests considerable scope for future research to develop and test interventions for mitigating confidence biases caused by feedback, especially in cases where the feedback is uninformative.

References

Alicke, M. D., & Govorun, O. (2005). The better-than-average effect. In *The Self in Social Judgment* (pp. 85–106).

Ambuehl, S., & Li, S. (2018). Belief updating and the demand for information. *Games and Economic Behavior*, *109*, 21–39.

Arkes, H. R., Christensen, C., Lai, C., & Blumer, C. (1987). Two methods of reducing overconfidence. *Organizational Behavior and Human Decision Processes*, *39*(1), 133–

[16] While some component of these effects may be explained as statistical artefacts (Benoit & Dubra, 2011; Krajc & Ortmann, 2008), the evidence suggests that a component remains after controlling for individual differences and in contexts using elicitation procedures that eliminate the statistical issues (Benoît et al., 2015; Burks, Carpenter, Goette, & Rustichini, 2013; Eil & Rao, 2011).

144.

Banerjee, R., Datta Gupta, N., & Villeval, M. C. (2020). Feedback Spillovers Across Tasks, Self-Confidence and Competitiveness. *SSRN Electronic Journal, Gate Working Paper No, 1806*. Retrieved from https://papers.ssrn.com/abstract=3157516

Barron, K. (2020). Belief updating: does the "good-news, bad-news" asymmetry extend to purely financial domains? *Experimental Economics*, 1–28.

Barron, K., & Gravert, C. (2018). Confidence and Career Choices: An Experiment. *SSRN Electronic Journal*.

Benoît, J., & Dubra, J. (2011). Apparent Overconfidence. *Econometrica*, *79*(5), 1591–1625.

Benoît, J., Dubra, J., & Moore, D. A. (2015). Does the Better than Average Effect Show that People are Overconfident: Two Experiments. *Journal of the European Economic Association*, *13*(2), 293–329.

Benoit, J. P., & Dubra, J. (2011). Apparent overconfidence. *Econometrica*, *79*(5), 1591–1625.

Berger, J., & Pope, D. (2011). Can Losing Lead to Winning? *Management Science*, *57*(5), 817–827.

Berlin, N., & Dargnies, M.-P. (2016). Gender differences in reactions to feedback and willingness to compete. *Journal of Economic Behavior & Organization*, *130*, 320–336.

Billett, M. T., & Qian, Y. (2008). Are Overconfident CEOs Born or Made? Evidence of Self-Attribution Bias from Frequent Acquirers. *Management Science*, *54*(6), 1037–1051.

Brookins, P., Lucas, A., & Ryvkin, D. (2014). Reducing within-group overconfidence through group identity and between-group confidence judgments. *Journal of Economic Psychology*, *44*, 1–12.

Burks, S. V., Carpenter, J. P., Goette, L., & Rustichini, A. (2013). Overconfidence and Social Signalling. *The Review of Economic Studies*, *80*(3), 949–983.

Buser, T., Gerhards, L., & van der Weele, J. (2018). Responsiveness to feedback as a personal trait. *Journal of Risk and Uncertainty*, *56*(2), 165–192.

Buser, T., Niederle, M., & Oosterbeek, H. (2014). Gender, Competitiveness, and Career Choices. *The Quarterly Journal of Economics*, *129*(3), 1409–1447.

Camerer, C., & Lovallo, D. (1999). Overconfidence and Excess Entry: An Experimental Approach. *American Economic Review*, *89*(1), 306–318.

Cárdenas, J.-C., Dreber, A., von Essen, E., & Ranehill, E. (2012). Gender differences in competitiveness and risk taking: Comparing children in Colombia and Sweden. *Journal of Economic Behavior & Organization*, *83*(1), 11–23.

Castagnetti, A., & Schmacker, R. (2020). Protecting the Ego: Motivated Information

Selection and Updating. *Working Paper*.

Chen, S., & Schildberg-Hörisch, H. (2019). Looking at the bright side: The motivational value of confidence. *European Economic Review*, *120*, 103302.

Coutts, A. (2019). Good news and bad news are still news: experimental evidence on belief updating. *Experimental Economics*, *22*, 369–395.

Daniel, K., & Hirshleifer, D. (2015). Overconfident Investors, Predictable Returns, and Excessive Trading. *Journal of Economic Perspectives*, *29*, 61–88.

Denning, J., Murphy, R., & Weinhardt, F. (2020). Class Rank and Long-Run Outcomes. *National Bureau of Economic Research, NBER Working Paper, No. 27468*. Retrieved from http://www.nber.org/papers/w27468.pdf

Dohmen, T., Falk, A., Huffman, D., & Sunde, U. (2010). Are Risk Aversion and Impatience Related to Cognitive Ability? *American Economic Review*, *100*(3), 1238–1260.

Drobner, C., & Goerg, S. J. (2021). *Optimistic belief updating and ex-post rationalization of ego-relevant information*. Working Paper.

Dunning, D. (2011). The dunning-kruger effect. On being ignorant of one's own ignorance. In *Advances in Experimental Social Psychology* (Vol. 44). Academic Press.

Eberlein, M., Ludwig, S., & Nafziger, J. (2011). The Effects of Feedback on Self-Assessment. *Bulletin of Economic Research*, *63*(2), 177–199.

Ederer, F. (2010). Feedback and Motivation in Dynamic Tournaments. *Journal of Economics & Management Strategy*, *19*(3), 733–769.

Eil, D., & Rao, J. M. (2011). The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself. *American Economic Journal: Microeconomics*, *3*(2), 114–138.

Eriksson, T., Poulsen, A., & Villeval, M. C. (2009). Feedback and incentives: Experimental evidence. *Labour Economics*, *16*(6), 679–688.

Ertac, S. (2011). Does self-relevance affect information processing? Experimental evidence on the response to performance and non-performance feedback. *Journal of Economic Behavior & Organization*, *80*(3), 532–545.

Fischbacher, U. (2007). z-Tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, *10*(2), 171–178.

Fischer, M., & Sliwka, D. (2018). Confidence in knowledge or confidence in the ability to learn: An experiment on the causal effects of beliefs on motivation. *Games and Economic Behavior*, *111*, 122–142.

Fischer, M., & Wagner, V. (2019). Efects of Timing and Reference Frame of Feedback

Evidence From a Field Experiment. *Rationality and Competition CRC TRR 190 Working Paper Series, No 150*. Retrieved from www.rationality-and-competition.de

Fishbach, A., Eyal, T., & Finkelstein, S. R. (2010). How Positive and Negative Feedback Motivate Goal Pursuit. *Social and Personality Psychology Compass*, *4*(8), 517–530.

Gervais, S., & Odean, T. (2001). Learning to Be Overconfident. *Review of Financial Studies*, *14*(1), 1–27.

Gill, D., & Prowse, V. (2012). A Structural Analysis of Disappointment Aversion in a Real Effort Competition. *American Economic Review*, *102*(1), 469–503.

Gotthard-Real, A. (2017). Desirability and information processing: An experimental study. *Economics Letters*, *152*, 96–99.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125.

Große, N. D., & Riener, G. (2010). Explaining gender differences in competitiveness: Gender-task stereotypes. *(No. 2010, 017). Jena Economic Research Papers.*

Grossman, Z., & Owens, D. (2012). An unlucky feeling: Overconfidence and noisy feedback. *Journal of Economic Behavior & Organization*, *84*(2), 510–524.

Günther, C., Ekinci, N. A., Schwieren, C., & Strobel, M. (2010). Women can't jump?—An experiment on competitive attitudes and stereotype threat. *Journal of Economic Behavior & Organization*, *75*(3), 395–401.

Hilary, G., & Menzly, L. (2006). Does Past Success Lead Analysts to Become Overconfident? *Management Science*, *52*(4), 489–500.

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

Hollard, G., Massoni, S., & Vergnaud, J.-C. (2016). In search of good probability assessors: an experimental comparison of elicitation rules for confidence judgments. *Theory and Decision*, *80*(3), 363–387.

Huang, L., & Murad, Z. (2020). Feedback Spillover Effect on Competitiveness Across Unrelated Tasks. *Behavioral Research in Accounting*, *32*(1), 69–85.

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

Kuhnen, C. M., & Tymula, A. (2012). Feedback, Self-Esteem, and Performance in Organizations. *Management Science*, *58*(1), 94–113.

Larrick, R. P., Burson, K. A., & Soll, J. B. (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.

Malueg, D. A., & Yates, A. J. (2010). Testing Contest Theory: Evidence from Best-of-Three Tennis Matches. *Review of Economics and Statistics*, *92*(3), 689–692.

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology*, *79*(3), 280–295.

Mezulis, A. H., Abramson, L. Y., Hyde, J. S., & Hankin, B. L. (2004). Is There a Universal Positivity Bias in Attributions? A Meta-Analytic Review of Individual, Developmental, and Cultural Differences in the Self-Serving Attributional Bias. *Psychological Bulletin*, *130*(5), 711–747.

Möbius, M., Niederle, M., Niehaus, P., & Rosenblat, T. (2014). Managing Self-Confidence. *NBER Working Paper*.

Moore, D. A. (2007). Not so above average after all: When people believe they are worse than average and its implications for theories of bias in social comparison. *Organizational Behavior and Human Decision Processes*, *102*(1), 42–58.

Moore, D. A., & Cain, D. M. (2007). Overconfidence and underconfidence: When and why people underestimate (and overestimate) the competition. *Organizational Behavior and Human Decision Processes*, *103*(2), 197–213.

Moore, D. A., & Healy, P. J. (2008). The Trouble With Overconfidence. *Psychological Review*, *115*(2), 502–517.

Moutsiana, C., Charpentier, C. J., Garrett, N., Cohen, M. X., & Sharot, T. (2015). Human Frontal-Subcortical Circuit and Asymmetric Belief Updating. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *35*(42), 14077–14085.

Murad, Z. (2016). Does group discussion lead to better informed and more strategic market entry decisions? *Economics Letters*, *144*, 25–28.

Murad, Z., Sefton, M., & Starmer, C. (2016). How do risk attitudes affect measured confidence? *Journal of Risk and Uncertainty*, *52*(1), 21–46.

Murphy, R., & Weinhardt, F. (2020). Top of the class: The importance of ordinal rank. *Review of Economic Studies*.

Petrusic, W. M., & Baranski, J. V. (1997). Context, Feedback, and the Calibration and Resolution of Confidence in Perceptual Judgments. *The American Journal of Psychology*, *110*(4), 543.

Pulford, B. D., & Colman, A. M. (1997). Overconfidence: Feedback and item difficulty effects. *Personality and Individual Differences*, *23*(1), 125–133.

Reuben, E., Sapienza, P., & Zingales, L. (2012). The emergence of male leadership in competitive environments. *Journal of Economic Behavior & Organization*, *83*(1), 111–117.

Rose, J. P., & Windschitl, P. D. (2008). How egocentrism and optimism change in response to feedback in repeated competitions. *Organizational Behavior and Human Decision Processes*, *105*(2), 201–220.

Rosenqvist, O., & Skans, O. N. (2015). Confidence enhanced performance? – The causal effects of success on future performance in professional golf tournaments. *Journal of Economic Behavior & Organization*, *117*, 281–295.

Ryvkin, D., Krajč, M., & Ortmann, A. (2012). Are the unskilled doomed to remain unaware? *Journal of Economic Psychology*, *33*(5), 1012–1031.

Sharot, T., Korn, C. W., & Dolan, R. J. (2011). How unrealistic optimism is maintained in the face of reality. *Nature Neuroscience*, *14*(11), 1475–1479.

Subbotin, V. (1996). Outcome Feedback Effects on Under- and Overconfident Judgments (General Knowledge Tasks). *Organizational Behavior and Human Decision Processes*, *66*(3), 268–276.

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

Wozniak, D., Harbaugh, W. T., & Mayr, U. (2014). The Menstrual Cycle and Performance Feedback Alter Gender Differences in Competitive Choices. *Journal of Labor Economics*, *32*(1), 161–198.

Zizzo, D. J. (2010). Experimenter demand effects in economic experiments. *Experimental Economics*, *13*, 75–98.

Appendix A1: The Circle Task

The Circle task has been used previously in Hollard et al., (2016) to study calibration between absolute confidence and success rates. The task involves seeing a pair of black circles with white dots in them for one second and judging which circle has more dots. We were specifically attracted to this task because we wanted to measure confidence in a skill task that is insensitive to effort. As a skill task, it requires a minimum effort level to achieve the maximum performance after which additional effort does not improve performance. This reduces the possibility of incentive effects caused by the feedback that may affect effort and performance in the task. We also aimed to use a gender-neutral task both in terms of actual performance and perceptions about the task. We pre-tested the task for these properties in a standard experimental session format with 30 subjects recruited via Orsee (Greiner 2004). The experiment was computerized and programmed with Ztree (Fischbacher 2007). The pre-test was conducted one month prior to the main study.

Four difficulty levels of a task were pretested: 50&51, 50&55, 50&60, 50&65, 50&70 dot circles presented in sets of 20 circle pairs per each difficulty level. We randomized the order in which the sets were presented. We also presented two of the difficulty levels (50&51 and 50&60) to the subjects three times, in order to check whether there were learning effects. We had two between-subject treatments, low and high incentives, to check whether incentives improve performance by promoting effort. The Low incentive scheme paid £0.05 per correct answer plus a fixed payment of £5. The high incentive scheme paid £0.50 per correct answer and no fixed payment. Subjects were paid for one randomly selected set.

We did not find a difference in performance between the two incentive levels across the difficulty levels. We did not find any difference in the answers of subjects to the questions "how hard did you try", and "how focused were you on a task" across incentive levels either, indicating that the subjects did not consciously vary effort depending on the rewards. The maximum average performance was observed in 50&70 dots set (19.81 in low and 19.75 in high incentive, Wilcoxon ranksum *p-value* = 0.729) and the minimum average performance was in 50&51 dot set (11.3 in low and 10.9 in high incentive, *p-value* = 0.437). We chose these two difficulty levels as our easy and difficult treatments for the main experiment. These two difficulty levels also showed the lowest variance in performance among subjects. We tested for learning effects and found that performance within a difficulty level did not improve from earlier to later sets (Cuzick, (1985) test for trend *p-value* = 0.432). There was significant

correlation between the actual score in the task and subjects' guess of their score (Figure A1) but more so for the easy (50&70 dot difference) than the difficult tasks (50&51 dot difference).

Previous studies have presented mixed evidence of whether task type being perceived to be "female" or "male" in nature can affect confidence levels between genders (Grosse & Riener 2010; Cardenas, Dreber, Essen & Ranehill 2011; Gunther, Ekinci, Schwieren & Strobel 2010). To avoid a possible gender effect in our study we checked whether there were differences in performance, response times, and perceptions about the difficulty level of the task across genders. To check for the perceptions about the task after each set of a task and at the end of the pre-test session we asked several questions (e.g. on a scale of 1 to 7, "how difficult did you find the task?", "how much did luck or skill determine your performance in the set?", "how hard did you try?"). We did not find any gender differences either in performance (*p-value* = 0.125), response times (*p-value* = 0.587) or in the perceptions of the task (*p-value* = 0.214). Table A2 reports the regression of average performance scores on observable characteristics of subjects.

Based on the pre-testing, we concluded that the Circle task is well suited to study confidence snowballing because:

i. incentives (and thus effort) play no role in actual performance,
ii. learning effects are absent,
iii. there are no gender differences
iv. performance and perceptions are not correlated with other measures of socio-economic variables.

Table A1: Performance at Each Difficulty and Incentive Level

Difficulty Level	**High Pay**	**Low Pay**	***p-value***
50&51	10.9	11.3	*0.437*
50&55	15.0	15.0	*0.941*
50&60	18.1	18.2	*0.767*
50&65	19.0	19.7	*0.119*
50&70	19.8	19.8	*0.729*
N	15	15	

p-values are from Wilcoxon ranksum test

Table A2: DV average performance score

Female	-0.86 (.77)
Age	-.12 (0.09)
Belief_LuckSkill	.50 (.311)
Risk_General	.15 (.21)
Confidence_General	.05 (.24)
Competitiveness	-.29 (.28)
Eyesight	-.08 (1.08)
Wearing Glasses	-1.19 (.91)
Constant	17.9 (0.30)***
N	22
R^2	0.643

** 10%, *** 1% significance levels. N=22 as some subjects chose not to report their gender.*

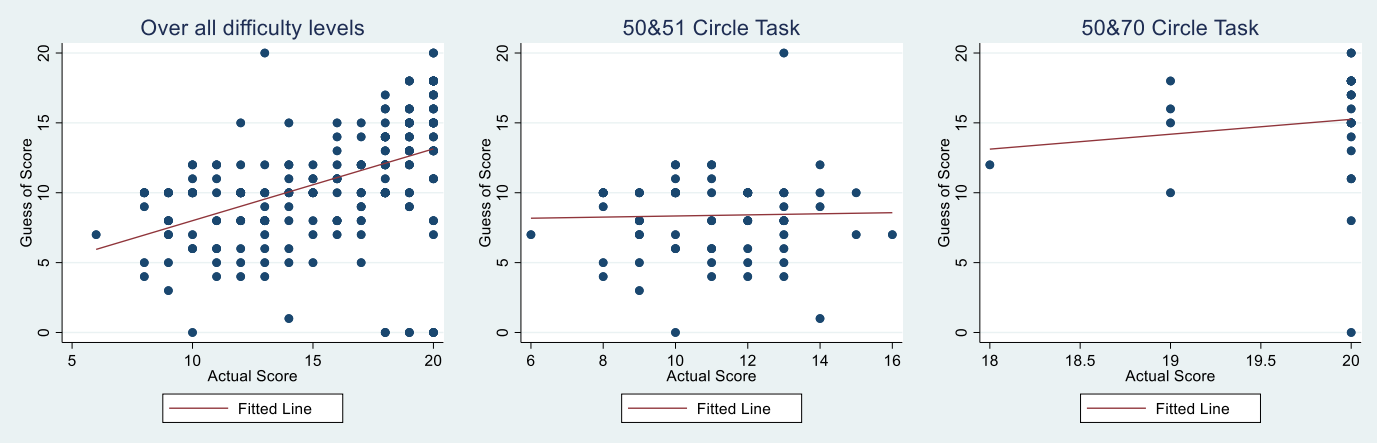


Figure A1: Scatter graph of guesses of score versus actual score in the pre-test

Appendix A2: Confidence Elicitation Tool

The elicitation tool (Figure 2 in the main text) is designed as a mechanism to elicit individual confidence (*Conf*). It has the attraction of being both simple to understand and incentive-compatible under both the Expected Utility and Rank Dependent Utility (RDU) models of choice under risk. Unlike some other incentivised methods for eliciting confidence (see Murad, Sefton, & Starmer, 2016), it does not rely on the assumption that individuals are risk-neutral. To see this consider the following analysis using the more general case of RDU where Options A and B refer to the options in Figure 2 and $u(.)$ and $w(.)$ are, respectively, utility and probability weighting functions. Applying RDU, the overall values of the options in the elicitation task of Figure 2 are given by:

$$RDU(Option\ A) = u(£10)w(Conf) + u(£3)(1 - w(Conf))$$

$$RDU(Option\ B) = u(£10)w(p) + u(£3)(1 - w(p))$$

We identify a *Probability Equivalence* (PE) as the value of p at an individual's switch point in any table. At that point, we assume $RDU(Option\ A) = RDU(Option\ B)$ – although this is only approximately true of course, given that we adjust p in the table in discrete steps. On this assumption, we may write:

$$u(£10)w(Conf) + u(£3)(1 - w(Conf)) = u(£10)w(PE) + u(£3)(1 - w(PE))$$

Rearranging gives,

$$u(£10)[w(Conf) - w(PE)] = u(£3)[\big(1 - w(Conf)\big) - \big(1 - w(PE)\big)]$$

Assuming that u and w are both increasing, it follows that $Conf = PE$.

Hence, we elicit $Conf$ from the switch point in a way that is independent of the specific forms of $u(.)$ and $w(.)$. We note that individual ambiguity attitudes could also affect the choice in our elicitation table. However ambiguity attitudes may be considered as conceptually related to confidence and even to the extent that it is a distinct phenomenon, as we are primarily interested in confidence snowballing across stages, some ambiguity-sensitive preference would not confound our main conclusions, so long as it is constant within-individual for the duration of the experiment.

The PE elicitation method is robust to Benoit & Dubra (2011) and Benoit, Dubra & Moore (2015) critique of apparent overconfidence. In their 2015 paper, they argue that belief and confidence elicitation methods should satisfy the following five conditions. There should be: (i) a well-defined scale that subjects are responding to behaviourally rather than with self-reports; (ii) a well-defined group they are comparing themselves to; (iii) a well-defined performance task they are asked to report their confidence about. They recommend elicitation of whole belief distributions about rankings (rather than expected rankings) (iv) and finally the propose that (v) elicitation should be appropriately incentivized. Our design satisfies all of these requirements. Similar arguments are also presented in Krajc and Ortmann (2008) where the authors show that unskilled and unaware (or Dunning-Kruger effect; Dunning, 2011) findings may be due to the noisy perception of one's own performance (which can happen in our setup but relative confidence measure is more robust to this), regression to the mean (again relative confidence measure is more robust to this), and a skewed distribution of skills in the group under consideration (we focus on confidence beliefs relative to the median rather than mean).

In our experiment, subjects fill in a Multiple Price List and choose lotteries with objective probabilities, satisfying condition (i). In relation to condition (ii), subjects always have a well-defined group they are comparing themselves to and receive full information on how groupings are determined in each stage (satisfying condition ii). Subjects carry out a well-defined performance task which they practice before the start of the payoff-relevant experiment (satisfying iii). By eliciting their confidence about scoring in the top half of their group we elicit their full subjective probability distributions since scoring in top and bottom halves are mutually exclusive events: subjects are then assumed to assign belief of *(100 - Conf)%* to scoring in the bottom half (satisfying iv). And finally, the belief elicitation is incentivized by an incentive-compatible mechanism which controls for risk attitudes, as explained above (satisfying v).

Appendix A3: The effect of the difficulty levels

Previous research has found differences between confidence levels in easy and difficult tasks (Moore, 2007; Moore & Cain, 2007). We find consistent evidence with this, where Stage 1 confidence is slightly higher in easy than in difficult variant of the Circle task. Looking at Table 1 in the main text and Table A3 below, we can see the level differences in confidence between the difficult and easy tasks to be around 4% points: this difference is significant when we look at the whole data combined, but not significant when we look at each treatment separately (Wilcoxon ranksum *p-value* = 0.203; 0.567 and 0.105 respectively in the NoFeedback, OwnFeedback and FullFeedback treatments). There is also no significant diff-in-diff of confidence levels between the treatment conditions and difficulty (the interaction term in Table A3). One reason for this can be that the task we selected was a neutral task that subjects had no experience in and no feeling of how easy or difficult the tasks is to start with both for them and others (confirmed by the pre-test of the task and discussed in Appendix A1). Hence subjects' confidence levels were quite calibrated (not significantly different from 50%) to start with. This is different from previous studies which use general knowledge quizzes, IQ tasks, math puzzles to elicit confidence levels (Ryvkin et al. 2012; Moore and Cain 2007; Coutts 2018; Berlin and Dargnies 2016 among others) and find significant differences between difficult and easy tasks. In that sense our task choice is justified that initial calibration responds to task difficulty minimally; hence our environment does start from ground 0 (uncontaminated by previous experiences and perceptions about the task) and we are able to track the confidence evolution from this ground 0.

In Table A4, we focus exclusively on updating of confidence behaviour after receiving feedback and whether the difficulty level of the task affects how confidence snowballs. We look at the data for feedback treatments (OwnFeedback and FullFeedback) and find that there are no significant differences in how subjects update their confidence in the easy and difficult variants of the task. This is not to say that there are no level differences in confidence levels between the difficulty levels: for example, for the group TopTop we observe subjects being 7.55% points more confident in the easy variant than in the difficult variant of the task (*p-value* = 0.062). This is not the case for the BottomBottom group however where 2.70% point difference is not statistically significant.

Table A3: Predicting Stage 1 Confidence: Interaction between difficulty level and treatments

OwnFeedback	0.95 (3.32)
FullFeedback	-1.15 (3.42)
Difficult	-4.91 (3.68)
OwnFeedback × *Diffiult*	3.26 (4.68)
FullFeedback × *Difficult*	0.69 (4.51)
Constant	44.91*** (12.18)
Controls	YES
N	384
R-Sq	0.0259

Controls include gender and nationality dummies, age, general willingness to take risk, general level of confidence and competitiveness and whether any issues with eyesight was reported.
** p < 0.10, *** p < 0.01 of the underlying coefficient Standard deviations are clustered in session level.*

Table A4: Testing confidence snowballing interacting with the task difficulty

	TopTop	BottomBottom	TopTop	BottomBottom
Stage 2	6.18***	-5.07**	7.77***	-4.31
	(1.69)	(2.18)	(2.18)	(3.97)
Stage 3	11.74***	-12.50***	14.30***	-7.92*
	(1.92)	(3.16)	(2.74)	(5.00)
Difficult	-7.55*	-2.70	-4.76	0.86
	(3.97)	(4.02)	(3.95)	(4.73)
Stage 2 × Difficult			-3.19	-1.52
			(3.38)	(4.38)
Stage 3 × Difficult			-5.14	-9.17
			(3.82)	(6.25)
Constant	22.69	55.17*	21.30	53.39*
	(21.75)	(31.43)	(21.63)	(32.00)
Controls	Yes	Yes	Yes	Yes
N	216	216	216	216
Adj R_sq	0.145	0.087	0.141	0.088

** 10%, ** 5%, *** 1% significance levels. Robust standard errors clustered at subject level in parentheses. FullFeedback and OwnFeedback treatments. Control variables include actual performance score, gender and nationality dummies, age, general willingness to take risk, general level of confidence and competitiveness and whether any issues with eyesight was reported.*

Appendix B:

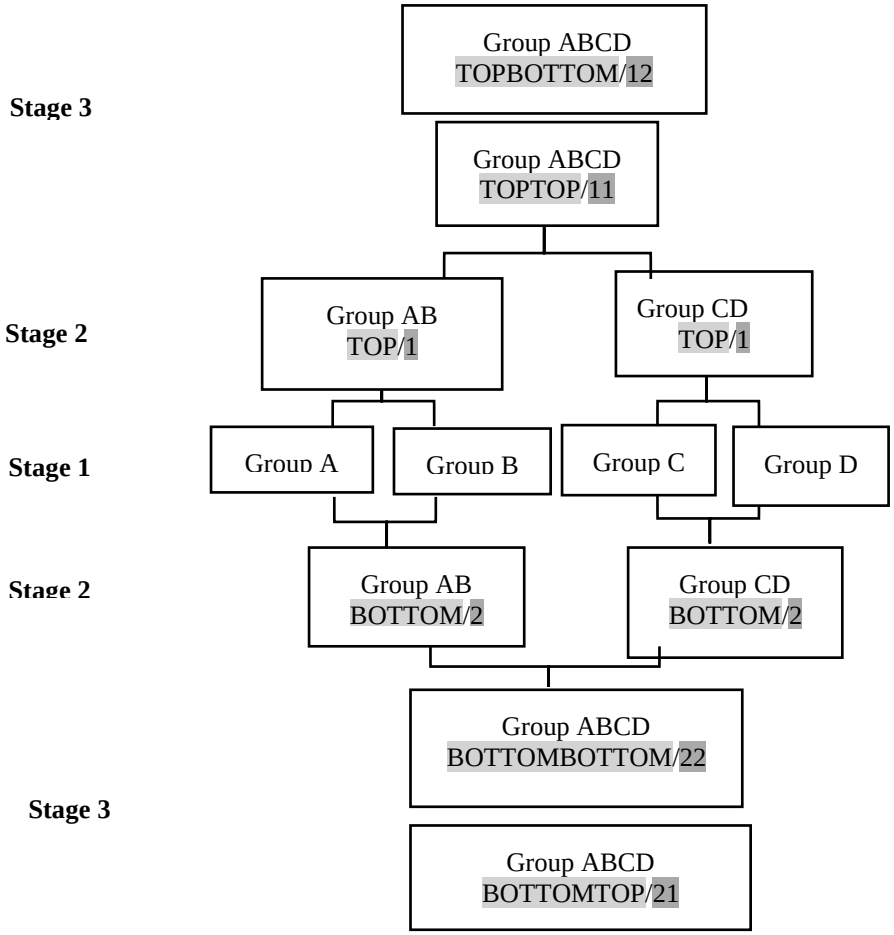


Figure B1: Matching at Each Stage of the Experiment

Subjects saw the words Top/Bottom in FullFeedback or the numbers 1/2 in NoFeedback and OwnFeedback treatments.

Instructions

Welcome!

You are about to participate in an experiment. There are 16 people participating in the experiment. You must not communicate with anyone. If you have a question at any time during the experiment, please raise your hand and someone will come to your desk to answer it.

The use of electronic devices such as mobile phones, music players, and tablet computers is strictly prohibited. Please make sure that all such devices are turned off and put away out of sight.

If you break these rules, you will be excluded from the experiment without receiving any payment and be disqualified from future experiments with CeDEx.

The Circle Task

In this experiment you will complete the Circle task. In this task you will see a pair of circles on your screen for 1 second. One of the circles contains 50 dots and the other contains 51 dots. After the circles disappear from your screen, you will be asked to judge whether the right or the left circle contained more dots. You have to indicate your judgement by pressing on the "Left" or "Right" button. When you press the button of your choice, you will move to the next pair of circles.

To acquaint you with the task, you can practice the task before we proceed with the remainder of the instructions. You will practice one set of 20 tasks. The practice set will not affect your final outcome. You can start the practice set by pressing the TASK 1 button.

[After everyone has finished the practice set:]

The experiment will consist of 3 stages. At each stage you will be in one of four groups consisting of four people. At each stage, you will have to do two things: fill out the Table and then complete the Circle Task.

Stage - Table:

For every row of the table you must make a choice between Option A and Option B. You should consider which of these two options you would prefer to have for each row then mark your choice by ticking the circle corresponding to your preferred option. In each table, we ask that you start with the top row and work your way down the rows. Option A will pay off £10 if you are one of the top two scorers of your group in this stage and £3 if you are one of the bottom two scorers of your group in this stage. (You will complete the Circle task and find out if you are in the top two or bottom two after you have filled the table.) Option B is a lottery which will pay off £10 or £3 with the chance of £10 progressively decreasing as you move down the rows. The chance of winning £10 from Option B in the first row is 100%. This is at least as good as the chance of being in the top two of your group, so we think that you will

want to choose Option B in the first row. But, since Option B gets progressively worse as you move down the rows (while Option A stays exactly the same), there may come a row where the chance of winning £10 of Option B is sufficiently small, that you prefer Option A. If you find such a row, you should then choose Option A for that row and the rows below it (since Option B continues to get worse all the way down the table).

Row	Option A:Bet	Your Choice:	Option B: Lottery
1		A ○ ● B	£10 with 100% chance
2		A ○ ○ B	£10 with 95% chance and £3 with 5% chance
3		A ○ ○ B	£10 with 90% chance and £3 with 10% chance
4		A ○ ○ B	£10 with 85% chance and £3 with 15% chance
5		A ○ ○ B	£10 with 80% chance and £3 with 20% chance
6		A ○ ○ B	£10 with 75% chance and £3 with 25% chance
7		A ○ ○ B	£10 with 70% chance and £3 with 30% chance
8		A ○ ○ B	£10 with 65% chance and £3 with 35% chance
9	You get £10 if	A ○ ○ B	£10 with 60% chance and £3 with 40% chance
10	you are one of the	A ○ ○ B	£10 with 55% chance and £3 with 45% chance
11	top two scorers of your group	A ○ ○ B	£10 with 50% chance and £3 with 50% chance
12	and you get £3 if	A ○ ○ B	£10 with 45% chance and £3 with 55% chance
13	you are one of the	A ○ ○ B	£10 with 40% chance and £3 with 60% chance
14	bottom two scorers of your group	A ○ ○ B	£10 with 35% chance and £3 with 65% chance
15		A ○ ○ B	£10 with 30% chance and £3 with 70% chance
16		A ○ ○ B	£10 with 25% chance and £3 with 75% chance
17		A ○ ○ B	£10 with 20% chance and £3 with 80% chance
18		A ○ ○ B	£10 with 15% chance and £3 with 85% chance
19		A ○ ○ B	£10 with 10% chance and £3 with 90% chance
20		A ○ ○ B	£10 with 5% chance and £3 with 95% chance

[Instructions for FullFeedback, NoFeedback and OwnFeedback treatments, differences highlighted in grey]

Stage - Circle Task

After you have completed the table you will complete the Circle task. You will have 20 tasks to complete similar to the practice set. After everyone has completed the task, your scores will be calculated. For each correct answer, 1 point will be added to your score so that you can score up to 20 points. [NoFeedback: Your score will be compared to the scores of the other members of your group and this will determine whether you are in the TOP or BOTTOM two of your group for that stage.] [FullFeedback and OwnFeedback: Your score will be compared to the scores of the other members of your group and at the end of the stage you will be informed of whether you were in the TOP or BOTTOM two of your group for that stage.] Ties will be randomly broken. You will then be assigned to a new group of four participants [FullFeedback: according to your performance] and the next stage will begin.] [NoFeedback: You will not learn if you were in the TOP or Bottom of your groups until the end of the experiment.]

Stages

There are 3 stages in this experiment. At each stage, all 16 participants will be divided into groups of 4. At each stage, you will complete the Circle task and receive a score based on your performance. Similarly, each other group member will complete the Circle task, seeing an identical set of circles, scoring 1 point for each correct answer, and so scoring up to 20 points. Before completing the Circle task of each stage you will fill out the table as described above. [FullFeedback and OwnFeedback: After each stage, you will get feedback whether you were in the top or bottom of your group.] The groupings at each stage are explained below.

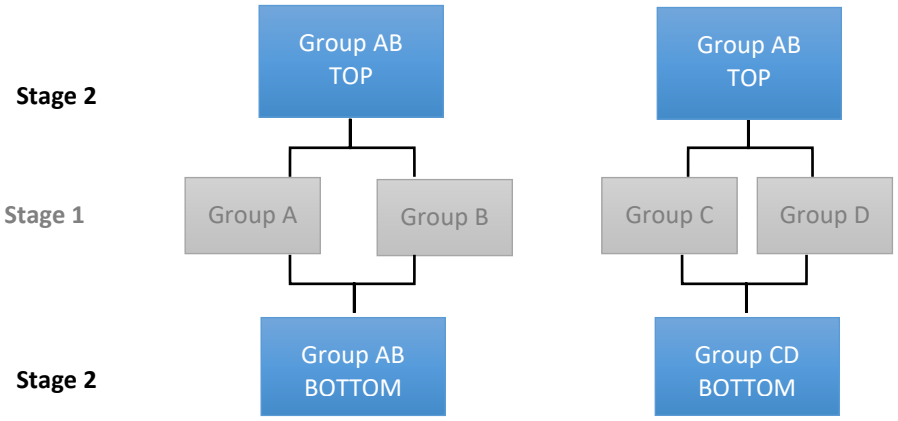
Stage 1:

In Stage 1, you will be randomly matched with 3 other participants to form a group. There will be four groups: Group A, Group B, Group C and Group D. You will see which group you are in on your screens.

[FullFeedback] Stage 2:

In Stage 2, you will be placed in a new group according to your scores in Stage 1.

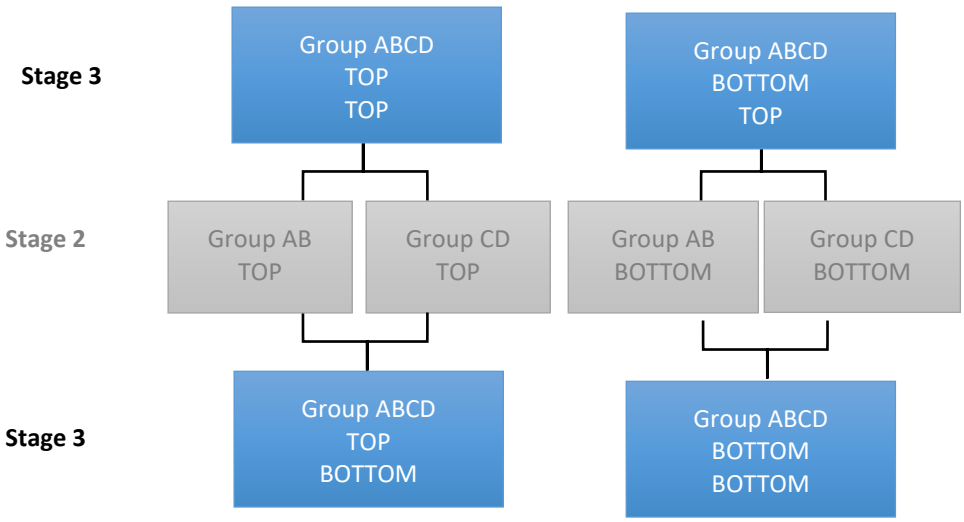
- The top two scorers of Group A will be matched with the top two scorers of Group B to form the **Group AB_Top**.
- The top two scorers of Group C will be matched with the top two scorers of Group D to form the **Group CD_Top**.
- The bottom two scorers of Group A will be matched with the bottom two scorers of Group B to form the **Group AB_Bottom**.
- The bottom two scorers of Group C will be matched with the bottom two scorers of Group D to form the **Group CD_Bottom**.



[FullFeedback] Stage 3:

In Stage 3, you will be placed in a new group according to your scores in Stage 2.

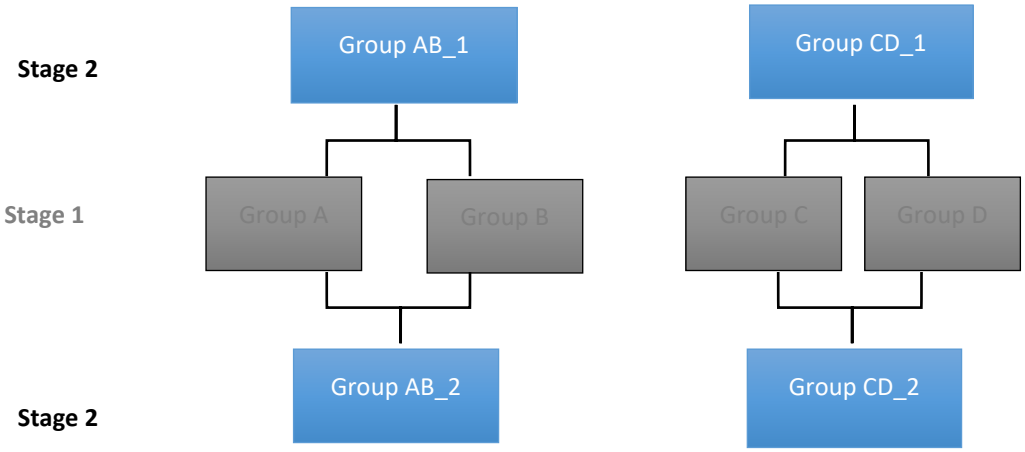
- The top two scorers of Group AB_Top will be matched with the top two scorers of Group CD_Top to form the **Group ABCD_TopTop**.
- The bottom two scorers of Group AB_Top will be matched with the bottom two scorers of Group CD_Top to form the **Group ABCD_TopBottom**.
- The top two scorers of Group AB_Bottom will be matched with the top two scorers of Group CD_Bottom to form the **Group ABCD_BottomTop**.
- The bottom two scorers of the Group AB_Bottom will be matched with the bottom two scorers of Group CD_Bottom to form the **Group ABCD_BottomBottom**.



[OwnFeedback and NoFeedback]: Stage 2:

In Stage 2, you will be placed in a new group

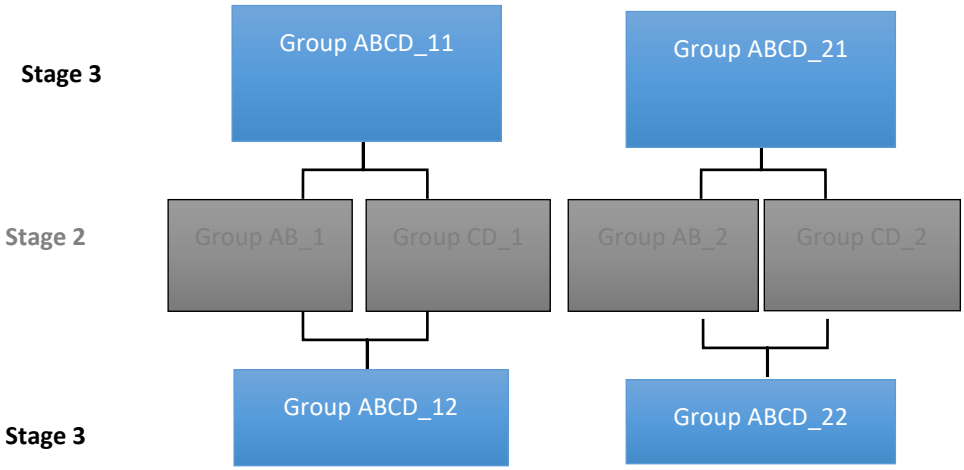
- Two participants of Group A will be matched with two other participants of Group B to form the **Group AB_1 and AB_2**.
- Two participants of Group C will be matched with two other participants of Group D to form the **Group CD_1 and CD_2**.



[OwnFeedback and NoFeedback]: Stage 3:

In Stage 3, you will be placed in a new group.

- Two participants of Group AB_1 will be matched with two other participants of Group CD_1 to form the **Group ABCD_11 and ABCD_12**.
- Two participants of Group AB_2 will be matched with two other participants of Group CD_2 to form the **Group ABCD_21 and ABCD_22**.



Payment

At the end of the session, a 6-sided die will be rolled. Depending on the outcome, you will be paid based on one of the stages and either the table or your performance in the Circle task.

Roll	Paid according to:
1	Stage 1 – Table
2	Stage 1 – Circle
3	Stage 2 – Table
4	Stage 2 – Circle
5	Stage 3 –Table
6	Stage 3 – Circle

If a table is selected, a 20-sided die will be rolled to select a row of that table and you will be paid according to your choices on the selected row. If you have chosen Option B on the selected row, you will play out the lottery of the selected row. If you have chosen Option A on the selected row, you will be paid £10 if you were in the top two of your group and £3 if you were in the bottom two of your group according to your scores in the selected stage.

If the Circle task is selected, you will get £0.50 for each correct answer so that you can earn up to £10.

This payment structure has been designed so that you have an incentive to do your best in the Circle task and to think carefully about each and every choice you have to make in the Table.

If you have read and understood the instructions you may start the experiment. If you have any questions, please raise your hand and someone will come to your desk to answer it.

Appendix C

	Instructions for the Circle Task: Practice the Circle Task		
	Instructions for all of the experiment		
Stage 1	**Groupings**	You are in Group A/B/C/D	
	Confidence	Fill in Confidence Elicitation Table	
	Task	Complete set of Circle Tasks	
	Feedback	"Press Continue to start the next Stage"	NoFeedback
		"Your Circle task Performance was in TOP/BOTTOM half of your group. Press Continue to start the next Stage"	OwnFeedback & FullFeedback
Stage 2	**Groupings**	You are in Group AB_1/AB_2/CD_1/CD_2	NoFeedback & OwnFeedback
		You are in Group AB_Top/AB_Bottom/CD_Top/CD_Bottom	FullFeedback
	Confidence	Fill in Confidence Elicitation Table	
	Task	Complete set of Circle Tasks	
	Feedback	"Press Continue to start the next Stage"	NoFeedback
		"Your Circle task Performance was in TOP/BOTTOM half of your group. Press Continue to start the next Stage"	OwnFeedback & FullFeedback
Stage 3	**Groupings**	You are in Group ABCD_11/ABCD_12/ BCD_21/ABCD_22	NoFeedback & OwnFeedback
		You are in Group ABCD_TopTop/ABCD_BottomBottom/ ABCD_TopBottom/ABCD_BottomTop	FullFeedback
	Confidence	Fill in Confidence Elicitation Table	
	Task	Complete set of Circle Tasks	
	Feedback	"Press Continue to start the next Stage"	NoFeedback
		"Your Circle task Performance was in TOP/BOTTOM half of your group. Press Continue to start the next Stage"	OwnFeedback & FullFeedback
		Post-Study Questionnaire	

Table C1: Experimental Flow

Table C2: Performance Statistics in the Circle Task (Stage 1 score)

	Easy Circle Task			Difficult Circle Task		
	Mean	St.dev	Median	Mean	St.dev	Median
TopTop	19.97	0.14	20	12.1	1.55	12
BottomBottom	18.92	2.93	19.5	8.94	1.82	9
TopBottom	20	0.01	20	11.71	1.43	12
BottomTop	19.5	0.65	20	8.86	2.13	9
Wilcoxon ranksum p (ToptTop=BottomBottom)		0.000		0.000		

N=48 per feedback condition and difficulty level.

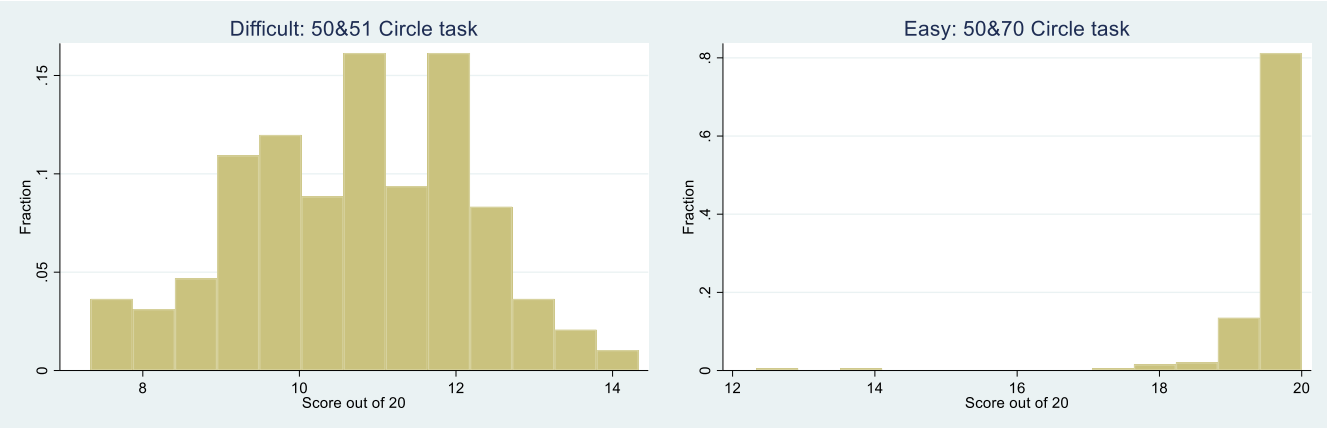


Figure C1: Distribution of average scores in the two difficulty variants of the task.

Table C3: Testing for snowballing differences between subjects that are confident versus not confident.

DV: Confidence	**TopTop**		**BottomBottom**	
	Model 1	Model 2	Model 1	Model 2
Stage 3	10.00*	32.50**	9.00	-13.00
	(5.05)	(10.95)	(13.66)	(7.73)
Confident	48.50***	44.75***	59.50***	35.00
	(8.65)	(8.75)	(4.75)	(1.61)
Stage3×Confident	-10.00*	-19.5	-34.00	-14.50*
	(4.94)	(13.57)	(20.50)	(7.73)
FullFeedback		2.32		-14.50**
		(10.48)		(6.47)
Stage 3×FullFeedback		22.50*		22.00
		(12.04)		(15.88)
Confident×FullFeedback		3.68		24.50***
		(12.26)		(5.08)
Stage 3×Confident×FullFeedback		9.50		-19.50
		(14.43)		(22.21)
Constant	28.60***	26.25***	20.50**	35.00***
	(5.98)	(8.65)	(6.17)	(1.61)
Adj R^2	0.695	0.635	0.466	0.493
N	24	42	30	44

** 10%, *** 1% significance level. Subjects classified as Confident=1 if first stage confidence is > 65% and Confident =0 if <45%. Standard errors clustered at session level.*