Eberlein, Marion; Ludwig, Sandra; Nafziger, Julia

Working Paper

Effects of Feedback on Self-Assessment

Bonn Econ Discussion Papers, No. 39/2005

Provided in Cooperation with:
Bonn Graduate School of Economics (BGSE), University of Bonn

Suggested Citation: Eberlein, Marion; Ludwig, Sandra; Nafziger, Julia (2005) : Effects of Feedback on Self-Assessment, Bonn Econ Discussion Papers, No. 39/2005, University of Bonn, Bonn Graduate School of Economics (BGSE), Bonn

This Version is available at:
https://hdl.handle.net/10419/22945

Bonn Econ Discussion Papers

Discussion Paper 39/2005

Effects of Feedback on Self-Assessment

by

Marion Eberlein, Sandra Ludwig, Julia Nafziger

December 2005

Bonn Graduate School of Economics
Department of Economics
University of Bonn
Adenauerallee 24 - 42
D-53113 Bonn

The Bonn Graduate School of Economics is
sponsored by the

Deutsche Post World Net
MAIL EXPRESS LOGISTICS FINANCE

The Effects of Feedback on Self-Assessment

Marion Eberlein, Sandra Ludwig, and Julia Nafziger*

University of Bonn, Adenauer Allee 24-42, 53113 Bonn, Germany.

marion.eberlein@uni-bonn.de, sandra.ludwig@uni-bonn.de, jnafziger@uni-bonn.de

December 21, 2005

Abstract

It is a well-known phenomenon that people tend to overestimate their relative abilities. Psychological studies show that a vast majority of people thinks that their ability is above the average when they have to assess their position in a distribution of a target group. We analyse in an experiment whether this is still true when people receive feedback on their relative abilities. Our main finding is that feedback influences agent's decisions and improves overall self-assessment.

Keywords Overconfidence, Feedback, Experimental Economics
JEL Classification: C91, D01, D80

*We would like to thank Alexander Koch, Matthias Kräkel and seminar participants at the Monday's Workshop in Bonn and the Brown Bag Seminar in Karlsruhe for helpful comments. Financial support from the Deutsche Forschungsgemeinschaft (DFG), grant KR 2077/2-1, is gratefully acknowledged. Part of this research was conducted, while Nafziger stayed at the IDEI and she would like to thank the IDEI and the University of Toulouse and for their great hospitality.

1 Introduction

It is a well-known phenomenon that individuals tend to overestimate their abilities, what is referred to as being overconfident (e.g. Fischhoff et al. (1977) and Lichtenstein and Fischhoff (1977)). The most famous example is that a majority of individuals think that they have better driving skills than the average driver (Svenson (1981)). However, in studies like this one it is not always clear, whether agents are really overconfident in the sense that their self-assessment is systematically biased (i.e. mistakes do not cancel out on average). While people have many occasions to assess their driving skill and that of others, many experiments involve tasks which are new to subjects and therefore lead to mistakes or a too high prior. One way to disentangle random mistakes from systematic biases is to let subjects repeatedly solve a task and give them feedback about their (relative) ability.[1] Were agents actually biased (and not only make mistakes or have a too high prior), repetition and feedback should not reduce overestimation, thus, there is overconfidence. Had they simply a too high prior about being better than the average, because they never solved such a task before, then for example, feedback and repetition should reduce overestimation, i.e. there is no real bias. The aim of our experiment is to examine the effects of feedback about the relative ability on individuals' self-assessments. We define overconfidence as the systematic overestimation of one's relative ability.[2]

The effects of feedback on overconfidence have already been examined in some psychological studies, where subjects perform calibration tasks with ambiguous results. While Pulford and Colman (1997) and Sharp et al. (1988) find that feedback does not influence overconfidence, Adams and Adams (1961) and Lichtenstein and Fischhoff (1977) find that feedback reduces overconfidence or can improve calibration, respectively. However, these studies differ in the form of feedback that is provided. Pulford and Colman (1997) give outcome feedback (the correct answers to the posed questions), whereas the others give performance feedback (information about realism of confidence) or statistical feedback (training in probability and calibration).

Our experiment differs in the following important dimensions from the psychological studies: we introduce performance related payments[3], subjects do not have to carry out a calibration task, but simply estimate their relative ability, the instructions are not "framed" and finally we try to classify different groups, who react differently to feedback. Using performance related payments provides subjects with the right incentives to try hard to make the correct self-assessment. This is important since some of the ambiguities in the experiments by psychologists might be due to the lack of such payments. The calibration task (stating confidence intervals for the answers to questions) is quite difficult and maybe a reason why feedback does not necessarily help if subjects are not familiar enough with it. Further, assessment of one's *relative* ability (compared to absolute ability which most psychologists considered) seems most

[1] There are other techniques to reduce overconfidence in decision making as the de-biasing technique used by Koriat et al. (1989) who let the subjects write down reasons why an answer to a question is correct or wrong.

[2] There are other definitions of overconfidence, e.g. that people overestimate their ability (without comparison to others), overestimate the precision of private relative to public signals, have an illusion of control or are miscalibrated in the sense that they overstate the precision of their own estimates.

[3] Lichtenstein and Fischhoff (1977) pay subjects, but the payment is not performance related.

relevant for economic settings: For example, in a tournament not the absolute ability matters, but the relative abilities of the player and his opponent(s). Additionally, it may even be easier to assess one's relative ability than the absolute one. Moreover, psychologists "framed" their instructions, which implies that people might be influenced in their decisions just by the wording of the instructions which makes the interpretation of results more difficult. We want to avoid this possible influence by using a neutral language.[4] Lastly, it is interesting to see how different individuals react to feedback, i.e. to classify different groups of individuals rather than just focus on the average impact of feedback. This helps to understand better what proportion of subjects is overconfident and especially why subjects are overconfident.
The aim of our study is to examine the effects of the economically most relevant form of feedback, circumventing the aforementioned problems. Subjects answer multiple choice questions over several rounds and then assess repeatedly their relative ability, which is simpler for subjects not used to statistics than a calibration task. They receive very precise feedback about their relative performance and their number of correctly answered questions (i.e. the absolute ability) before a new round starts. This provides hence indirect information on the goodness of calibration (concerning relative ability) and direct information on the absolute and relative ability. The feedback about one's relative and absolute ability corresponds rather to outcome feedback (like giving the correct answers to questions) than to performance feedback and may facilitate learning. We pay subjects (the amount depends on the decisions) and use a neutral language in the instructions: we call e.g. the "better than the average group" "group A" and so on.

The most closely related papers in economics are the experiment by Camerer and Lovallo (1999) and the study by Ferraro (2005). Camerer and Lovallo (1999) consider a market entry game in which individual payoffs depend on a subject's relative ability and find excess entry. Ferraro (2005) tests whether people assess their own abilities correctly or rather overestimate them. In the field experiment he conducts students are asked after an exam to estimate how many questions they answered correctly and their ranking. This procedure was repeated in the following two exams. Feedback in the form of personal grades and the distribution of grades in the first and second exam, respectively, showed little impact on self-estimation and overconfidence. Compared to our study subjects could not perfectly deduce from this information their relative rank. Moreover, since during the exams some time passed, subjects had the opportunity to prepare for the next exam. This preparation might influence self-assessment and could not be disentangled from the effects of feedback.
Our study is also related to the behavioral finance literature and the experiments conducted in this field (For an overview of this literature see Barberis and Thaler (2002)). Among these are so called cascade experiments, which test (among other things) whether people correctly apply Bayes' rule (e.g. Anderson and Holt (1997) who find consistent behavior, or Nöth and Weber (2003) who find overconfidence). Other experiments are interested in the implications of overconfidence on trading outcomes in the market. Financial market experiments ask (e.g.

[4]In some psychological studies also tricky questions are used which we try to avoid.

Biais et al. (2005)), whether people have a bias for private – compared to public signals, which is called overconfidence. It is typically observed that due to overconfidence there is too much trading.

The design of our experiment is roughly as follows. In the beginning the participants answer 90 questions, which are divided into six blocks. After each block of questions the subjects make a decision, for which the payoff depends on the relative ability of each subject compared to a control group. Thus, we can back out of the subjects' decisions their beliefs about their rank relative to the subjects in the control group. After each round – i.e. before making the decision for the current round – subjects are told, whether they performed better or worse than the median in answering the questions the round before. Lastly, subjects assess their overall ability in the end.

Our main finding is that feedback has some impact on the decisions of individuals, however not on all, thus there are different groups of individuals. Roughly one group of subjects ignores its feedback more or less completely, in another one subjects are from the beginning on relatively sure about their ability and some individuals seem to be unsure about their ability and react to the feedback. For the latter ones it is interesting to note that feedback does not always improve the decision, but seems to confuse some subjects, who make more mistakes in the later rounds than in the beginning. This is an interesting new finding, since from previous studies a wrong decision was often interpreted as a refusal to learn from feedback. We show that mistakes are not always due to ignorance, but to an overreaction to feedback. However, there are also subjects, for whom feedback improves the decisions. When assessing their overall ability almost all subjects do so correctly, which means that they used the feedback during the rounds to update their priors. We further examine, whether subjects are more likely to follow positive or negative feedback with the result that most subjects react more to good news.

The paper is structured as follows. Section 2 describes the experimental procedure. Then we present and discuss the results in Section 3 and the last Section concludes.

2 Experimental Design

The computerized experiment was conducted at the University of Bonn and programmed with z-Tree by Fischbacher (1999). A total of 30 students participated in two sessions for which subjects were recruited via the internet by using the ORSEE software by Greiner (2003). During the experiment, the subjects earned Taler, converted into Euros in the end, where 210 Taler = 1 Euro. Average hourly earnings were 10 Euros. The instructions[5] were read out loudly before the experiment started. Subjects also answered control questions to make sure that they understood the experimental procedure.

[5]Avaible upon request.

In the baseline question session (*Session Q*) 15 subjects answer 90 general knowledge (multiple choice) questions, which are divided into 6 blocks. After each block subjects state how many questions they think, they answered correctly in this block. The number of correctly answered questions will be called absolute ability or type of a subject in the following.
In the feedback session (*Session F*) 15 subjects answer the same questions, divided in the same 6 blocks. Here, subjects state after each block of questions whether they think that they belong to the upper or lower 50% compared to the subjects in *Session Q* in this block. The upper or lower 50% are determined as follows: Subjects are ranked according to their number of correctly answered questions, a higher number implying a higher rank. If two or more subjects have the same number of correctly answered questions, the subject who answered the correct questions faster receives the higher rank. The lower 50% are then subjects with ranks 1-8 and the upper 50% the ones with ranks 9-16. Thus, the relative ability (type) of an individual is, whether it belongs to the upper or lower 50%. Note that, we avoid formulations like "the best 50% of subjects" and call the upper and lower 50%, group A and B, respectively. Starting with the second block of questions, subjects receive the information to which group they belonged to – i.e. their relative type – and how many questions they answered correctly – i.e. their type – in the previous block. By telling the subjects how many questions the have correct we avoided one additional source of uncertainty. After their decision regarding block 6, subjects finally assess to which group they belong to when *all* questions are considered.

The payoffs for the questions and decisions are as follows. To avoid confounding hedging effects we randomly selected one of the six blocks for the question part and one for the decision part at the end of the experiment (by two independent draws). For each correctly answered question in this block subjects earned 270 Taler minus 0.9 Taler for every second a subject needed for its correct answers. The assessment task was rewarded based on the other randomly drawn block at the end of the experiment. In *Session Q* subjects received 300 Taler if they stated their number of correct answers correctly. In *Session F* a subject received 1500 Taler (roughly 7 Euros) if it placed itself in the correct group A or B respectively, otherwise it got 180 Taler (roughly 0.85 Euros). For a correct overall assessment after the last round the subjects got 300 Taler and for a wrong one 20. Finally, subjects in *Session F* got 525 Taler for showing up.

3 Results

Analysing *Session Q*, we find that the average type for each block of questions is 6.5 whereas the average belief of the subjects about their type is 8. This difference is not for every block significant (Wilcoxon test). Nevertheless, a difference of one and a half correct questions on average is quite large. If we divide subjects into those who are over- or underconfident[6] or those

[6]We define an overconfident individual as one that believes it has answered more questions correctly/is in a better group than it actually has/is. Overall we consider subjects as over-/underconfident, when in at least four blocks they over- or underestimate their type. The relatively unbiased subjects have three correct guesses.

who are unbiased, we see that a majority of subjects (73.3%) is overconfident. The average absolute difference between true type and belief – i.e. the average absolute bias – of the overconfident subjects is 3.27, of the underconfident subjects (13.3%) it is 1.72 and of the unbiased (13.3%) it is 0.79. This shows that people have severe problems with their self-assessment. Across rounds, these numbers stay roughly the same.[7] The average type in *Session F* is 6.6, thus almost the same as in *Session Q*.[8] Considering the relative self-assessment one sees that roughly half of the subjects are wrong each period (see below for more details) and that of this half the majority is overconfident.

Considering the overall effect of feedback on self-assessment, we note that feedback helps to improve the correctness of the predicted ability of the whole population.[9] In the last three decision rounds only 6 people on average (only 5 in the last) made the wrong decision. The average in the first three rounds is 7.7. However, there are differences between subjects. If we divide subjects into two groups – one group with all subjects, who change their belief over the rounds more than one time (Group II) and the other group with those, who change it zero or one time (Group I) one sees the following in Figure 1: The number of mistakes in Group II is actually increasing and for Group I it is decreasing. This finding could be interpreted in the following way: subjects, who are relatively unsure (in the sense that they change their belief very often) about their ability seem to get more confused by the feedback and make more and more mistakes after receiving more and more information. For subjects, who are relatively sure about their ability, feedback helps to improve the decisions further.

[Figure 1]

Are mistakes due to ignorance of feedback as one might suspect? In Figure 2, we see that a majority chooses the same action as they got feedback about the previous round. We divided the subjects, who react to feedback in two groups in Figure 3: In Group A are those subjects, who made less than two mistakes and in Group B are the ones, who made at least two mistakes in total. One sees that the group that makes more mistakes is more often in line with its feedback. This indicates that subjects, who make many mistakes do not ignore their feedback, but follow it "too" much, i.e. they overreact.

[Figure 2]

[Figure 3]

However, although some subjects seem not to profit from feedback, one should note that even though subjects make mistakes, they do not make the same mistake twice in a row. Without the one subject, who ignored feedback completely and made every round the wrong decision, in only 15 percent of all decisions the same mistake was made twice in a row, indicating that

[7] Types as well as beliefs do not differ significantly over rounds according to a Wilcoxon test.

[8] There is no significant difference across treatments according to a Mann-Whitney U test ($p > 0.5$).

[9] We do not find a significant improvement between the first round and the last one according to a Fisher exact test ($p = 0.355$ one-sided).

subjects seem to react to their feedback and do not repeat their mistakes. However, their reaction might be wrong.

Examining further the reaction to feedback, we see that subjects are more likely to follow "good news" which is shown in Figure 4: A majority of subjects chooses group A, although they received feedback B.[10] But interestingly, the second largest group is the one, that received feedback B (being in the lower 50 percent group) and followed this feedback. Thus, there is no clear evidence in our experiment that subjects ignore bad news as psychologists claimed.

If one looks at the percentage of subjects making mistakes for the groups in Figure 4 one can see that the largest fraction of mistake-making subjects can be found under those subjects, who ignore their feedback and here especially those subjects make the most mistakes, who claim to be in the upper 50% after receiving the opposite feedback.

[Figure 4]

In Figure 5 we can further see how feedback that is (not) constant (not) helps to improve the decision of subjects. We see here that subjects, for which the group (A or B) changed from the previous round to the actual round ("change in position"), make more often mistakes than subjects, whose position did not change.

[Figure 5]

We observe that the feedback about the relative type is not consistent across rounds for all subjects. Note that we consider feedback as consistent, when is changes at most once over time. This happens especially for those subjects where the type varies a lot over the different blocks, e.g. from 3 up to 11, and for those who are just around the median type When we account for this fact and look at subjects with consistent feedback separately, we see that in each round, the latter subjects represent on average 80.5% of those who make a correct decision. Interestingly, in the first decision (where subjects have not yet received any feedback) these subjects represent only 37.5% of those being correct. This indicates that feedback strictly improves their decision. According to a Fisher exact test, we can reject the hypothesis that correct and wrong guesses for subjects with consistent feedback are not related between the first and the last round in favor of the alternative that there are more wrong (correct) guesses in the first (last) round ($p = 0.0594$ one-sided). For subjects who have alternating feedback, we cannot reject the independence hypothesis if we compare average guesses of the first five rounds ($p = 0.28$ one-sided) with the last one (if we consider the first and last round, there is even the hint that guesses worsen ($p = 0.051$ one-sided)).

For the overall assessment in the end, feedback is not ignored. Here, only four people were wrong. One of them always said he belongs to the top 50 percent, although he was *always* told he does not. Two subjects got three times a good feedback and three times a bad one – and seemed to be thus unsure; one seemed to ignore its failures. Remarkably, ignoring failures did not occur more often, as it is observed by psychologists (See e.g. Barberis and Thaler (2002) for an overview). This might be due to the fact that in our experiment subjects are paid more

[10]However, it is still the larger fraction that followed the feedback.

for a correct decision.

For the final decision, observe again that those subjects who receive consistent feedback build the majority (73%) of those who make the correct guess. According to a Fisher exact test ($p = 0.0594$ one-sided), we can again reject the hypothesis that correct and wrong guesses for those subjects are not related between the first and the overall desision in favor of the alternative that there are more wrong (correct) guesses in the first (overall) decision. For subjects who do not have consistent feedback, we again cannot reject the independence hypothesis ($p = 0.5$ one-sided). Moreover, we see that subjects with higher abilities made the better final assessment – even if all subjects learned about their relative position quite well. Total types range from 26 to 56. Subjects with types larger than 41 correctly state that they belong to the top 50% (with one exception), whereas not all of those with types smaller than or equal to 39 state that they belong to the worst 50%.

Finally we want to have a closer look at subjects' different reaction to feedback: The first group, with 40% of the subjects, almost (i.e. at most one exception) has a constant feedback and also makes most times (with at most one exception) the right decision. Thus, this group consists of people, who are good or bad in the question task and also know or believe this as they made the right decision already in the first round. Their behavior is consistent with their feedback perhaps because their belief about themselves is reinforced by the feedback they receive. Out of this group, one subject always received the "good" feedback, except for one time. For the round, in which it got the "bad" feedback, it said that it belongs to the worse group. One of these subjects has always been told that it belongs to the lower 50% and it only said once that it belongs to the upper 50%. Lastly, one subject got once a "bad" feedback, which it ignored. The second group with 13% ignores feedback completely – incidentially these were also the people, who made most often mistakes when evaluating themselves, never getting it right or only once. These people seem to have no sense of their relative ability, which may be due to their unresponsiveness to feedback that they receive about their type. The third group (47%) is responsive to their feedback. One subject always followed the direction of the feedback, even though the feedback is about the relative ability for the questions of the block before. Five subjects ignored bad news in the beginning, but then, with some delay, after receiving repeatedly bad news, they started to change their choice. Two of those subjects changed their belief then so dramatically that they ignored good news in the end. Finally, one person ignored good news initially and followed more the bad news. Taking these observations together, one sees that feedback has a strong impact – on average only four people deviated from their feedback in each decision round.

4 Conclusion

In general, subjects tend to be biased (especially overconfident) when estimating their relative abilities. In our experiment we provide subjects with precise feedback about their relative ability. We find that this feedback is used differently by individuals. A large group clearly uses it to make their decisions, a small group ignores it and another group seems to know

their relative ability. Overall, feedback helps some subjects to make better decisions and thus to reduce mistakes, while others seem to get confused and make even more mistakes. These people do not make mistakes because they ignore their feedback, but because they do not ignore it.

Thus, overall it is not clear, whether "overconfidence" is a real bias or just a mistake, since the decrease in mistakes is driven by one group that uses the feedback in the right way, while for the others the mistake rate increases. Therefore, overestimation does not vanish completely over time.

References

[1] Adams, J, Adams P., 1961. Realism of Confidence Judgements. Psychological Review 68 (1), 33-45.

[2] Anderson, L, Holt, C. 1997. Informational Cascades in the Laboratory. American Economic Review 87 (5), 847-862.

[3] Barberis, N, Thaler, R., 2002. A Survey on Behavioral Finance. NBER Working Paper Nr. 9222.

[4] Biais, B, Hilton, D, Mazurier, K, Pouget, S., 2005. Judgemental Overconfidence, Self-Monitoring and Trading Performance in an Experimental Financial Market. Review of Economic Studies 72 (2), 287-312.

[5] Camerer, C, Lovallo, D., 1999. Overconfidence and Excess Entry: An Experimental Approach. American Economic Review 89 (1), 306-318.

[6] Ferraro, P.J., 2005. Know Thyself: Incompetence and Overconfidence. Experimental Laboratory Working Paper Series 2003-001, Georgia State University, http://epp.gsu.edu/pferraro/docs/OverconfidenceWorkingPaper2003-001revisedJanuary252005.pdf.

[7] Fischbacher, U., 1999. Z-Tree. Toolbox for Readymade Economic Experiments. IEW Working Paper 21, University of Zurich, http://www.iew.unizh.ch/ztree/index.php.

[8] Fischhoff, B, Slovic, P, Lichtenstein, S., 1977. Knowing With Certainty: The Appropiateness of Extreme Confidence. Journal of Experimental Psychology: Human Perception and Performance 3, 552-564.

[9] Greiner, B., 2003. The Online Recruitment System ORSEE - A Guide for the Organization of Experiments in Economics. Papers on Strategic Interaction, Max Planck Institute for Research into Economic Systems, Discussion Paper 10-2003, http://www.orsee.org/.

[10] Koriat, A, Lichtenstein, S, Fischhoff, B., 1980. Reasons for Confidence. Journal of Experimental Pschology: Human Learning and memory 6 (2), 107-118.

[11] Lichtenstein, S, Fischhoff, B., 1977. Do Those Who Know More Also Know More About How Much They Know? The Calibration of Probability Judgements. Organizational Behavior and Human Performance 20, 159-183.

[12] Nöth, M, Weber, M., 2003. Information Aggregation with Random Ordering: Cascades and Overconfidence. Economic Journal 113, 166-186

[13] Pulford, B, Colman, A., 1997. Overconfidence: Feedback and Item Difficulty Effects. Personality and Individual Differences 23 (1), 125-133.

[14] Sharp, G, Cutler, B, Penrod, S., 1988. Performance Feedback Improves the Resolution of Confidence Judgements. Organizational Behavior and Human Decision Processes 42, 271-283.

[15] Svenson, O., 1981. Are we all less risky and more skillful than our fellow drivers? Acta Psychologica 47, 143-148.

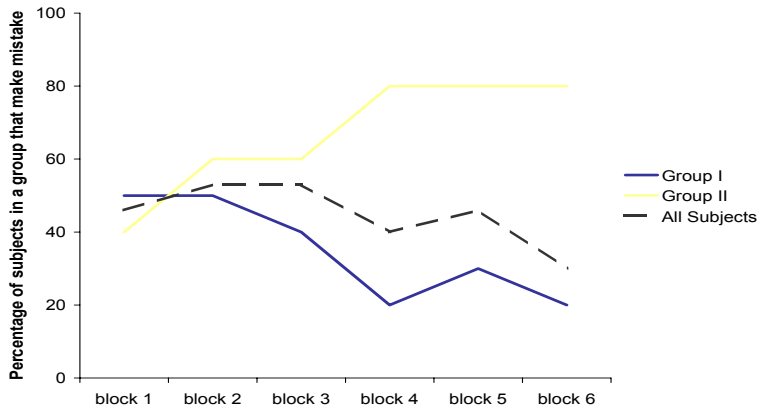

Figure 1: Mistakes over Rounds. Group I: subjects changing their belief less than or exactly one time; Group II: subjects changing their belief more than one time

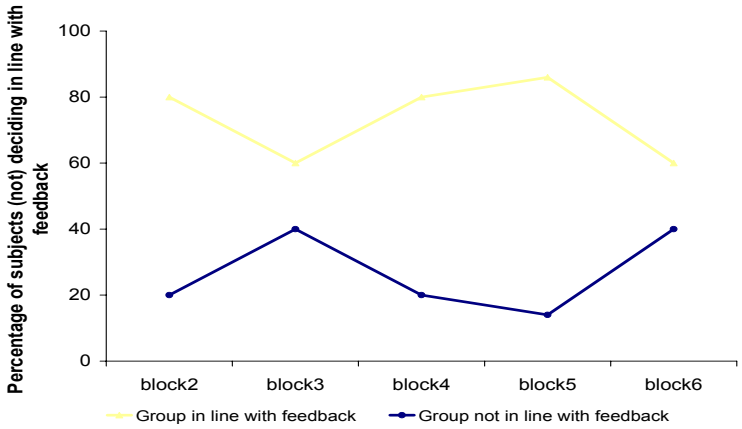

Figure 2: Percentage Reacting to Feedback

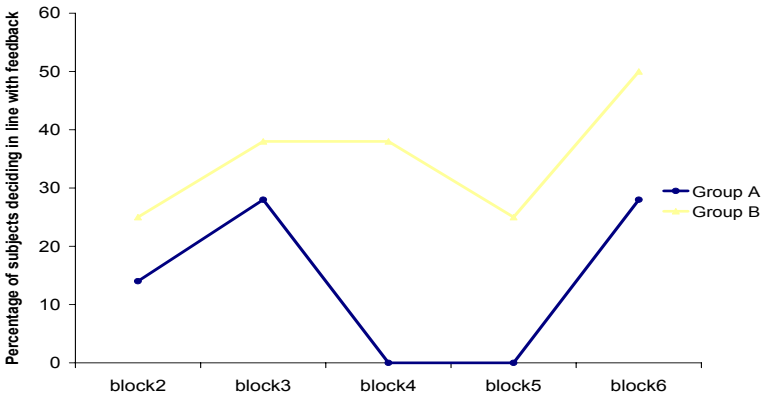

Figure 3: Reaction to Feedback. Group A: subjects that made less than two mistakes; Group B: subjects that made at least two mistakes.

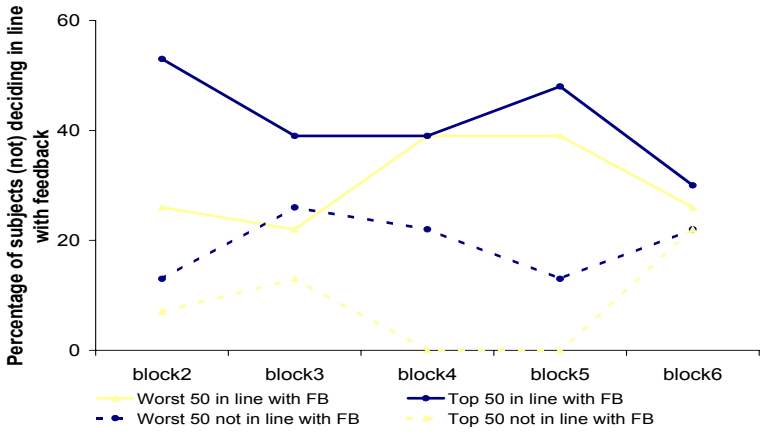

Figure 4: Reaction to Feedback 2. Top 50 and Worst 50 refers to the received feedback (A vs. B).

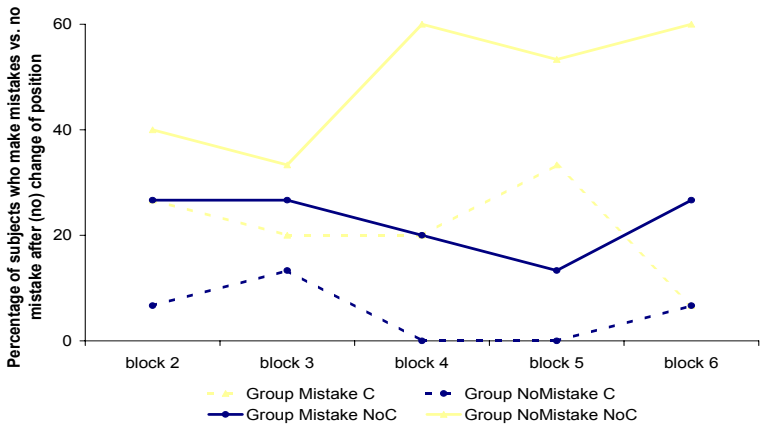

Figure 5: Subjects Making a Mistake or NoMistake After a Change (C) or No Change (NoC) in Position