To Know or Not to Know:
When does Feedback Lead To Better Assessment of Uncertainty of Own Beliefs?

Tanja M. Gruschke (tanjag@simula.no)
Simula Research Laboratory

Magne Jørgensen (magnej@simula.no)
Simula Research Laboratory

Abstract

People are frequently overconfident about the accuracy of their own beliefs. The goal of this experiment is to examine whether, and if so under what conditions, very large amounts of feedback lead to better assessments of the uncertainty of one's own beliefs. Fifteen participants answered the same 960 general knowledge questions over two days. Questions were selected from the board game "*Who wants to be a millionaire?*™". Each question had four answer alternatives. When an answer alternative had been chosen, the participants were asked to assess the probability that it was the correct answer. They did this by choosing from a list of predefined confidence intervals. The questions belonged to one out of six difficulty categories, as decided by the board game developers. Participants answered piles of 80 questions of the same difficulty at a time. Feedback about the correct answer was received immediately, and feedback about correspondence between "hit rate" and confidence level was received immediately after a pile of questions had been answered. In addition, a summary of the first day's performance was provided at the start of Day 2. We found that thirteen out of the fifteen participants improved in the correspondence between hit rate and confidence level on the second day, which suggests that the feedback had an effect. The strongest improvement was achieved on questions with high "global" difficulty, i.e., high difficulty as assessed by the board game developers, and low "internal" difficulty, i.e., low difficulty as perceived by the participants.

1. Introduction

A consistent finding in previous research on people's judgments regarding subjective probability is that their judgments are overconfident (Gilovich, Griffin, & Kahneman, 2002). In some areas, this overconfidence has been shown to be reduced after issuing timely and consistent feedback, e.g. in weather forecasting (Pliske, Crandall, & Klein, 2004) and in bridge playing (Keren & Teigen, 2001). In studies that make use of general-knowledge questions, significant reductions in overconfidence have as far as we know, not been achieved successfully with feedback (Keren & Teigen, 2001; Lichtenstein, Fischhoff, & Phillips, 1982; Subbotin, 1996; Zakay, 1992). It is, however, possible that the feedback issued in many of these studies has been such that did not enable people to improve their uncertainty assessments, e.g., the feedback may have been too sparse and infrequent. In Lichtenstein and Fischhoff (1980), for example, the participants did not receive information about the correctness of an individual question immediately after answering it, but about the proportion correct of a set of questions at a later stage. In this paper, we investigate the role of feedback when improving realism of uncertainty assessments in a learning-friendly environment, i.e., in a situation with many repetitions of question-answer-uncertainty assessment-feedback, combined with aggregated feedback when a set of questions has been answered.

To acquire a better understanding of when people are able to learn from feedback, it may be useful to know whether the degree of difficulty of the tasks affects learning. Previous research on judgment of subjective probability indicates that task difficulty does play a role regarding the accuracy of one's own beliefs (Griffin & Tversky, 1992; Lichtenstein & Fischhoff, 1980). Typically, easy uncertainty assessment tasks lead to underconfidence and hard tasks to overconfidence. In this study, we investigate whether there is a difference in degree of learning that depends on task difficulty.

In what follows we describe the study design (Section 2), present the results (Section 3) and summarize (Section 4).

2. Design of Study

2.1. Research Questions

Our research questions are as follows:

RQ1: To what extent do people improve their uncertainty assessment of correctness over answers to general knowledge questions when provided with immediate feedback and long learning sequences?

RQ2: What is the relation between question difficulty and uncertainty assessment improvement on answers to general knowledge questions?

2.2. Participants

Participants were recruited through different kinds of advertisements at the University of Oslo, i.e. notices posted on notice boards, information given in a software engineering class, and application of several mailing lists. We advertised for people who would enjoy answering a high number of general knowledge questions. Fifteen participants were recruited, all of whom expressed a pronounced motivation to participate. A high motivation was important, to ensure that the attentiveness towards answering the relatively high number of questions used in the experiment were of sufficient standard. The participants were paid 1000NOK (1 NOK = approx. 8 EUR).

2.3. Tasks

The participants were asked to answer general knowledge questions selected from the Norwegian version of the board game "Who wants to be a millionaire? ™". For each question, a participant chose one of the four available answer alternatives. When an answer was chosen, the participants were to assess how probable it was that they had selected the correct alternative by choosing from a set of seven predefined confidence levels. The participants were informed that the purpose of the experiment was to study the improvement of their assessments of probability (i.e. their uncertainty assessments.) The experiment lasted two days, with 4 hours of work each day.

The board game contains question cards. Each question card has a question with four alternative answers on each side and the correct answer printed on the opposite side. The questions are of various levels of difficulty, which level is indicated by the monetary value printed on the question cards. There are 15 difficulty categories, ranging from easy (labelled

1.000 NOK) to very difficult (labelled 2.000.000 NOK). In our experiment we used questions from the six difficulty categories marked with 10.000, 20.000, 40.000, 60.000, 80.000 and 100.000 NOK. The other difficulty categories were not used, because we believed them to be too easy or too difficult for the purpose of our experiment. Questions belonging to the same difficulty category and applied in a sequence in our experiment is, in this paper, defined as a "question pile". Each question pile applied contained 80 question cards (160 questions). The participants answered half the questions from six question piles each of the two days, i.e., they answered in total 80 * 6 * 2 = 960 questions. The questions were about, e.g.,, history, geography, politics, science, literature, sports, music, and cinema.

2.4. Measures

As described earlier, the questions had four alternatives answer. Therefore, the probability of choosing the correct alternative ranges from 0.25 (a random choice of alternative) to 1.00 (absolutely certain that the correct alternative has been chosen). We divided the probability range into seven intervals, each representing an uncertainty level; see Table 1. The participants chose one of these intervals when assessing the probability that a given answer alternative was the correct one. Each uncertainty level was complemented by a natural language description. A pilot study had indicated that all uncertainty levels would be used by the participants.

Table 1: Uncertainty Levels (see definitions below)

Uncertainty Level n	Level Range (Lev_n)	Natural language Description (Translated From Norwegian)
Level 0	[0.25]	"No idea"
Level 1	(0.25, 0.40]	Low confidence
Level 2	(0.40, 0.60]	Fifty-fifty
Level 3	(0.60, 0.75]	Good confidence
Level 4	(0.75, 0.90]	High confidence
Level 5	(0.90, 0.99)	Very high confidence
Level 6	[0.99, 1.00]	Absolute certainty

Definitions used in Table 1 and in the following analyses:

$_{min}Lev_n$ = minimum of probability interval n, e.g., $_{min}Lev_{1 =} 0.25$

$_{max}Lev_n$ = maximum of probability interval n, e.g., $_{max}Lev_{1 =} 0.40$

Lev_n = range of probability values between $_{min}Lev_n$ and $_{max}Lev_n$

$QLev_n$ = the number of questions with probability assessed to be within Lev_n.

$HitLev_n$ = the number of correct answers with probability assessed to be within Lev_n.

$HitRateLev_n = HitLev_n / QLev_n$ = proportion of correct answers of the total answers with probability assessed to be within Lev_n, referred to as the "hit rate".

Notice that Level 0 has no interval. The hit rate of that uncertainty level should be close to 0.25.

Ideally, the participants' $HitRateLev_n$ should be between $_{min}Lev_n$ and $_{max}Lev_n$. There has been an improvement in uncertainty assessment when $HitRateLev_n$ is outside $_{min}Lev_n$ and $_{max}Lev_n$ on the first day and inside on the second day. A change of $HitRateLev_n$ in the right direction is also considered improvement.

2.5. The Experiment

The experiment was conducted in one of the Department of Informatics' computer laboratories. Participants used standard installed internet browsers to access the web support system employed in the experiment. They were not allowed to use sources of information other than their own knowledge to answer the questions.

When the participants arrived on their first day, they were given a written description of the experiment and informed individually about the experimental process. They were also informed that a high correspondence between the hit rates and uncertainty levels should be their main goal. They were given the opportunity to discuss the natural language labels assigned to each uncertainty level; see Table 1. We ensured that the participants understood and accepted that an improved correspondence between hit rate and uncertainty level was the intention of the experiment before handing out the questions. To further motivate the participants to focus on improving their uncertainty assessments, instead of, for example, on getting as a high proportion of correct answers as possible, we informed them that the three participants who showed the most improvement in their uncertainty assessments would receive a prize.

The sequence of the question piles was random, i.e. the participants answered the question piles in different sequences from each other. However, the participants did answer the question piles in the same sequence on Days 1 and 2.

After receiving a question pile, the participants were informed about the difficulty category (printed on each card), e.g., that the questions in the question pile belonged to the 60.000 NOK difficulty category. Then, they started answering the questions. For each question, they chose an answer alternative and then assessed the confidence they had in their answer by choosing one of the uncertainty levels. Figure 1 illustrates the design of the input page. When an answer had been given and the level of uncertainty assessed, the participants checked and registered the correctness of their answer on the web-based system. When all the questions in a pile had been answered, they received feedback on the relation between the uncertainty levels (confidence levels) and the hit rates; see Figure 2 for an illustration. The feedback informed about the number of questions answered, the number of correct answers, and the hit rate for each level. The experimenter in charge helped with the interpretation of the feedback and gave support on how to improve the uncertainty assessments. The experimenter sat with the participants and went through all the uncertainty levels, commenting on the correspondence between them and the hit rates. The participants were encouraged to improve their uncertainty assessments through comments like: "*It seems as if you know more than you think, perhaps some of the answers you assessed as belonging to uncertainty level 1 really belongs to level 2*", or, "*You are a bit overoptimistic when assessing the uncertainty to be on levels 4 and 5.*". After the feedback had been examined and analyzed together with the experimenter, the participant was handed the next pile of questions. At the beginning of Day 2 the participants received a summary of their results from Day 1, which contained an overview of the correspondence between uncertainty levels and hit rates for each difficulty level. This summary was supplemented with an oral description and comments by the experimenter.

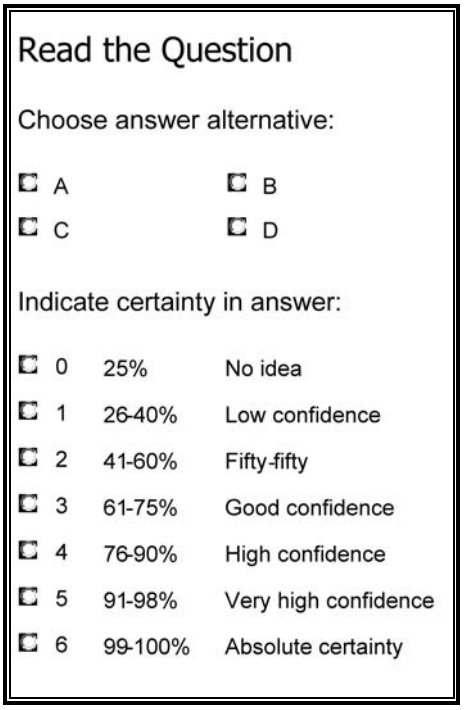

Figure 1: Input page for answer alternative and uncertainty assessment for each question.

Here you can see youre hit rates in relation to your level of confidence.

Contact the person in charge of the experiment before proceeding.

Confidence level		No. of answers	No. of correct answers	Hit rate (%)
0	25%	11	3	27. 27
1	26 -40%	19	7	36.84
2	41 -60%	12	8	66.67
3	61 -75%	14	10	71.43
4	76 -90%	9	9	100.00
5	91 -98%	3	3	100.00
6	99 -100%	16	16	100.00

Figure 2: Example of feedback page (translated from Norwegian) displayed immediately after a pile of questions were answered.

2.6. Threats to Validity

The participants themselves were responsible for checking and registering the correctness of their answer to a question. Since the participants were competing for a prize (lottery tickets), it is possible that cheating might have occurred. For example, the participants might have adjusted the correctness of their answer to achieve perfectly calibrated uncertainty levels. However, certain factors speak against the realization of this possibility. When talking with the participants, we found that those who admitted that they initially tried to keep score for their hit rate stated that they quickly lost interest in this rather difficult task. In addition, we believe that the potential reward was too small to outweigh the effort, and feeling, of cheating.

Another threat to the validity of the results is that the high number of questions to be answered could lead to boredom and indifference. Our observations of the participants suggest that this was not the case. The level of commitment and interest expressed by several of the participants during the whole experiment was, in fact, very high. Curiosity induced by the questions, and pleasure in displaying otherwise uncalled-for knowledge (such as specialized sports and celebrity facts) were frequently expressed by the participants. Spontaneous statements such as "*This is fun!*" make it safe to assume that the number of questions in itself did not generate much boredom and indifference. However, the large number of questions to be answered could, nevertheless, have led to a drop in concentration over the course of the experiment and hence reduced the possibilities for learning and improvement. To avoid this, the participants were allowed to take as many breaks as they desired, with each break being as long as they liked. They were, however, urged to take pauses in between, rather than in the middle of, a pile of questions.

3. Results

To examine whether the participants improved their uncertainty assessments we investigated the change in correspondence between hit rate and uncertainty level from Day 1 to Day 2. This analysis was conducted both with respect to the different levels of uncertainty (Section 3.1) and with respect to different levels of question difficulty (Section 3.2). In our analysis, a participant is assessed as improving his/her uncertainty assessments if there was a better correspondence between hit rate and uncertainty level on Day 2 than on Day 1.

3.1. Improvement from Day 1 to Day 2

Of the fifteen participants in the experiment, six achieved a clearly better correspondence between hit rate and uncertainty level on Day 2. Seven participants showed some improvement, while two participants did not show any improvement at all. Three had a good correspondence between hit rate and uncertainty level on Day 1 and maintained this performance on Day 2. The performance of each of the participants is displayed in Table 2. As can be seen from Table 2, the uncertainty levels where the participants had the most problems with hit rate correspondence, particularly on Day 1, were levels 3 and 4, i.e., uncertainty levels reflecting confidence in the range 61-90%..

Table 2: Individual Hit Rates (Hit rates within the corresponding probability intervals (Lev_n) are in bold letters)

Participant no.	Day	Lev_0	Lev_1	Lev_2	Lev_3	Lev_4	Lev_5	Lev_6
1	one	0.43	0.41	**0.44**	0.41	0.59	0.80	0.97
	two	0.35	0.55	0.73	**0.69**	0.70	0.86	**0.99**
2	one	**0.25**	**0.36**	**0.46**	**0.67**	**0.81**	**0.94**	**1.00**
	two	**0.27**	0.47	0.61	0.81	**0.82**	**0.92**	0.98
3	one	**0.29**	0.47	0.73	0.87	0.93	1.00	0.97
	two	**0.27**	0.45	**0.55**	0.82	**0.85**	**0.92**	**1.00**
4	one	0.36	**0.38**	**0.44**	**0.67**	**0.81**	0.83	0.97
	two	0.44	**0.33**	**0.44**	**0.64**	**0.82**	**0.95**	**1.00**
5	one	0.45	0.53	**0.55**	**0.64**	**0.83**	**0.95**	**1.00**
	two	0.49	0.54	**0.52**	0.78	**0.83**	1.00	**1.00**
6	one	0.37	0.43	**0.44**	**0.68**	0.56	0.86	0.98
	two	**0.25**	0.42	**0.49**	**0.72**	**0.86**	**0.91**	**1.00**
7	one	**0.30**	0.47	**0.47**	**0.63**	0.71	**0.93**	**1.00**
	two	**0.25**	**0.32**	**0.46**	**0.65**	0.97	**0.93**	0.98
8	one	**0.27**	0.41	**0.49**	**0.63**	**0.88**	1.00	**1.00**
	two	**0.23**	**0.30**	0.37	0.55	0.72	1.00	0.97
9	one	0.33	0.41	**0.53**	**0.60**	0.71	1.00	0.97
	two	0.33	**0.39**	**0.40**	0.53	**0.76**	0.89	**0.99**
10	one	**0.30**	**0.38**	**0.54**	0.83	**0.79**	1.00	**1.00**
	two	0.35	0.51	**0.50**	0.77	**0.75**	**0.90**	0.98
11	one	0.36	**0.39**	**0.60**	0.83	0.93	1.00	**1.00**
	two	0.37	0.67	0.63	0.83	**0.89**	1.00	**1.00**
12	one	0.32	0.41	**0.46**	**0.68**	**0.77**	0.77	0.98
	two	**0.24**	0.43	**0.56**	0.78	**0.83**	1.00	**1.00**
13	one	0.37	0.49	**0.54**	**0.63**	**0.81**	0.88	0.98
	two	0.39	**0.40**	**0.58**	**0.74**	**0.87**	**0.92**	0.98
14	one	0.36	0.48	0.62	0.59	0.67	0.87	**0.99**

	two	0.43	**0.40**	**0.51**	**0.67**	**0.78**	**0.95**	**0.99**
15	one	**0.28**	**0.34**	**0.44**	0.43	0.60	0.85	0.98
	two	**0.28**	0.45	0.36	0.46	0.51	**0.91**	**1.00**

3.2. Impact of Question Difficulty on Improvement of Uncertainty Assessment

There are two different types of question difficulty relevant to this context, i.e. the "global" and the "internal" question difficulty. The global difficulty is the difficulty as assessed by the board game provider. This refers to how difficult the board game providers believed that most individuals would find the task, i.e., with more difficult questions being those that fewer people would be able to answer correctly. To test whether the board game provider's assessments were valid for our participants, we examined the proportion of correct answers per difficulty category. As described earlier, the difficulty category was indicated by the amount of money printed on the question cards. Table 3 shows that the overall proportion of correct answers for all participants did decrease as the "global" question difficulty increased, which offers some degree of confirmation that the board game provider's assessment of "global" difficulty was correct.

Table 3: Hit Rate per Difficulty Category

Difficulty Category	Hit rate
10,000 NOK	0.73
20,000 NOK	0.68
40,000 NOK	0.59
60,000 NOK	0.57
80,000 NOK	0.55
100,000 NOK	0.51

"Internal" question difficulty, on the other hand, refers to how difficult a particular individual finds a question. In our experiment, this was measured as the uncertainty level (Lev_n), i.e., the level of confidence the individual has in the correctness of the answer.

Figures 3 and 4 display the correspondence between the average hit rate (y-axis) and the average uncertainty level (x-axis) for the "globally" hard questions (the 80,000 and 100,000 NOK questions) and for the "globally" easy questions (the 10,000 and 20,000 NOK questions). The shaded areas indicate the Lev_n, i.e. the level range of probability values between the minimum and maximum for that level. If a data point is inside the shaded area, this indicates a good correspondence between hit rate and level of uncertainty. The "ideal" hit rate is the mean value of the probability interval. Notice that we have, somewhat arbitrarily, set the accepted hit rate interval for Level 0 (the "no idea" category) to be [0.20-0.30].

Figures 3 and 4 suggest that the correspondence was already quite good on Day 1 for the "globally" hard questions (the 80,000 and 100,000 NOK questions) with four out of seven hit rates inside the probability intervals on Day 1, and with all but one hit rates inside on Day 2. Regarding the "globally" easy questions (the 10,000 and 20,000 NOK), there is not much improvement from Day 1 to Day 2. However, it is worth noting that, except for the uncertainty levels 0 and 1, the hit rates were inside the probability intervals on both days. This initially good correspondence, of course, makes it difficult to improve much.

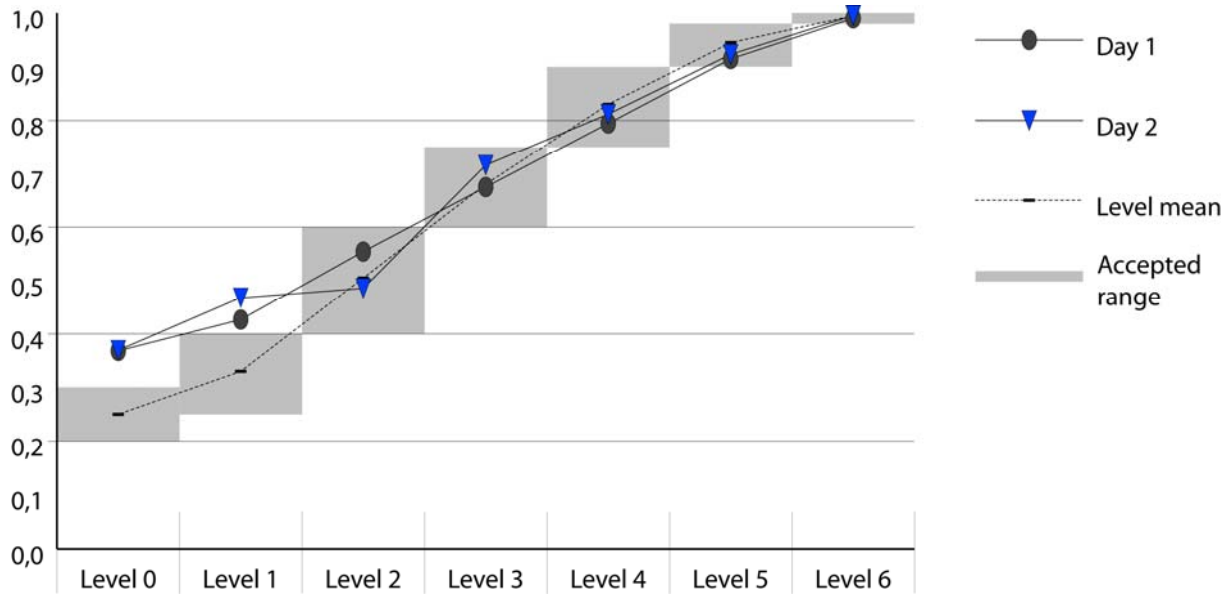

Figure 3: Hit Rates of "Globally" Hard Questions (80,000 and 100,000 NOK)

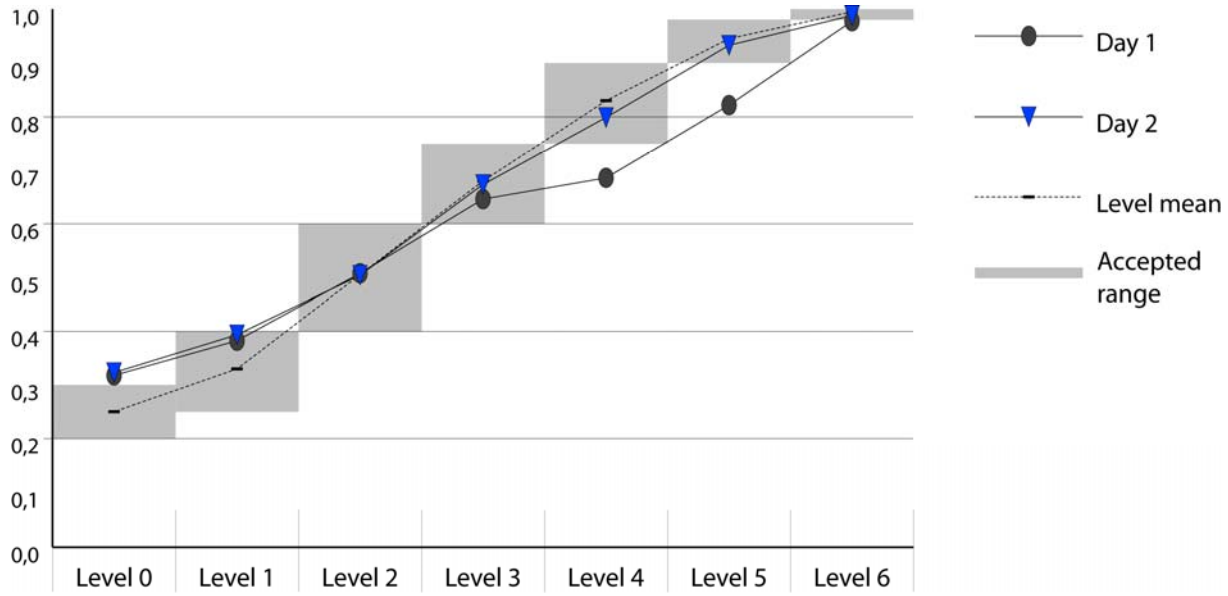

Figure 4: Hit Rates of "Globally" Easy Questions (10,000 and 20,000NOK)

We investigated each participant's performance on the uncertainty levels 0 and 1, i.e., the "internally" hard questions, and the uncertainty levels 5 and 6, i.e., the "internally" easy questions, for all question piles. We observed more improvement on the "internally" easy questions than on the "internally" hard questions. When looking at the "internally" easy questions, 11 out of the 15 participants had improved their uncertainty assessments, while only six of them improved on the "internally" hard questions. Figure 5 shows the correspondence between the hit rate and the uncertainty level for all "global" difficulty categories. From the

figure we see that there is a weak overconfidence on “internally” easy questions, i.e., uncertainty levels 5 and 6, which is reduced on Day 2. Regarding the “internally” hard questions, i.e., uncertainty levels 0 and 1, there is a weak underconfidence that does not change much on Day 2.

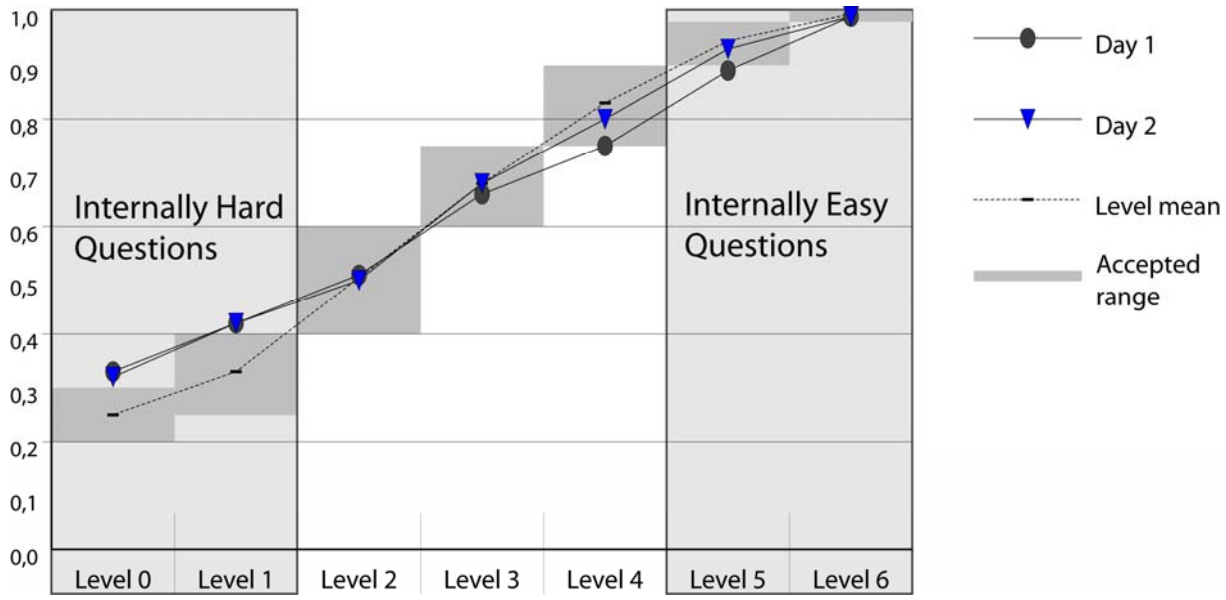

Figure 5: Hit Rate vs. of All Internal “Global” Difficulty Categories

4. Summary

In our study we investigated whether people’s ability to assess uncertainty improved when given long learning sequences and proper feedback. The main result was that thirteen out of the fifteen participants assessed the uncertainty better on the second day, which shows that the feedback provided had a positive impact on accuracy when assessing the uncertainty of one’s own beliefs. The strongest improvement was achieved when the “global” difficulty, i.e., the difficulty as perceived by the board game providers, was high and the “internal” difficulty, i.e., difficulty as perceived by the participants, was low. When the “global” difficulty was low, the participants had accurate uncertainty assessments initially and it was, consequently, difficult to improve. There was no improvement on questions with high “internal” difficulty, i.e., the uncertainty assessments were weakly underconfident even after extensive feedback. Our results differ, to some extent, with previous, more negative, results on people’s ability to improve uncertainty assessments of their own knowledge through feedback. The more positive results in our study may be a consequence of the more learning-friendly environment with immediate and extensive feedback, and long learning sequences.

References

Gilovich, T., Griffin, D., & Kahneman, D. (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge, United Kingdom: Cambridge University Press.

Griffin, D., & Tversky, A. (1992). The weighing of evidence and the determinants of confidence. *Cognitive Psychology, 24*, 411-435.

Keren, G., & Teigen, K. H. (2001). Why is p =.90 better than p =.70? Preference for definitive predictions by lay consumers of probability judgments. *Psychonomic Bulletin and Review, 8*(2), 191-202.

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior and Human Performance, 26*(2), 149-171.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Pliske, R. M., Crandall, B., & Klein, G. (2004). Competence in weather forecasting. In K. Smith, J. Shanteau & P. Johnson (Eds.), *Psychological investigations of competence in decision making* (pp. 40-68). Cambridge: Cambridge University Press.

Subbotin, V. (1996). Outcome feedback effects on under- and overconfident judgments (general knowledge tasks). *Organizational Behavior and Human Decision Processes, 66*(3), 268-276.

Zakay, D. (1992). The influence of computerized feedback on overconfidence in knowledge. *Behaviour & information technology, 11*(6), 329.