Error Rates and Uncertainty Reduction in Rule Discovery

M. Emrah Aktunc, Ceren Hazar, Emre Baytimur

forthcoming in

Review of Philosophy and Psychology

doi: 10.1007/s13164-020-00494-6

Three new versions of Wason’s 2-4-6 rule discovery task incorporating error rates or feedback of uncertainty reduction, inspired by the error-statistical account in philosophy of science, were employed. In experiments 1 and 2, participants were instructed that some experimenter feedback would be erroneous (control was original 2-4-6 without error). The results showed that performance was impaired when there was probabilistic error. In experiment 3, participants were given uncertainty reduction feedback as they generated different number triples and the negative effects of probabilistic error were not observed. These findings are informative not only about rule discovery tasks in general but also about contexts of inference under uncertainty.

Rule discovery is a general skill that is of central importance in both research on reasoning as well as studies on scientific cognition (Holyoak & Morrison, 2012; Sternberg & Ben-Zeev, 2001; Carruthers, Stich, Siegal, 2002; Feist, 2006). Wason’s well-known 2-4-6 task (1960) is one of the most studied rule discovery problems several different versions of which have been and are still being used as experimental paradigms (for a review, please see Evans, 2014).

In Wason's task, participants are given the number triple 2-4-6 and told that this triple obeys a certain rule set by the experimenter. They are asked to find out what this rule is by creating new number triples; for each triple the participant gets feedback from the experimenter on whether or not it obeys the rule. The participant may make guesses about the rule throughout the experiment and the experiment ends when the participant makes a final guess or when a certain amount of time runs out. The rule set by the experimenter is “ascending numbers.”

Wason's original task was informed and inspired by Popper's falsificationist philosophy of science (Popper, 1934; 1959). As the initial cue, 2-4-6 is very likely to strongly prime the participants to infer the rule is “ascending even numbers,” they go on trying new but similar triples, such as 8-10-12 or 24-26-28 etc. As a result, they get into a loop of getting positive feedback from the experimenter and most of them conclude that indeed the rule must be ascending even numbers, when in fact it is “any ascending numbers.” According to Wason, following Popper, these participants keep doing confirmatory tests, i.e. generate triples of ascending even numbers, that obey the hypothesis they already accepted to be true. However, had the participants tried to falsify their hypothesis, e.g. testing a triple of odd numbers such as 1-3-5, they would have still gotten the positive feedback from the experimenter and they could fairly easily have inferred that the rule is not about even numbers. By doing other types of falsificatory tests, e.g. trying 1-4-7 or 6-5-3, they would have been able to eliminate other possible hypotheses about the rule and eventually they would have discovered the correct rule. Indeed, as Wason reported (1960), the participants who successfully discovered the rule did precisely this, i.e. they tested triples such as 3-5-7 and 6-5-2 in addition to those like 8-10-12.

Wason’s 2-4-6 task is commonly seen as a behavioral task simulating scientific reasoning where the participants are like scientists devising and doing new experiments as they try different number triples. As one of the most well-known accounts in philosophy of science, it is no surprise that Popper's falsificationism does explain why many participants could not successfully discover the correct rule, but it does not say much about under what conditions participants may start doing more effective triple tests, i.e. testing triples such as 1-3-5 or 9-7-6 etc. One thing that may be useful is to make the participants aware that despite the positive feedback they get from the experimenter, there still is uncertainty about what the correct rule is. Indeed, this is true for scientific inference in general; sets of data may agree with one's hypothesis but, as history of science has taught us, the hypothesis may still be false. As an experimental task, Wason's 2-4-6 very effectively captures this fundamental aspect of scientific inference. Although Wason was inspired by Popper, who rejected the use of induction in empirical science, 2-4-6 has been adopted by some researchers as a valid experimental analog of inductive inference in the cognitive psychology of science. Hence, it would be useful if we consult a philosophical account of science that clearly formulates reliable inductive inference in devising new versions of 2-4-6 in which participants can more easily see the necessity and effectiveness of testing diverse triples. We aimed to make the 2-4-6 task more externally valid by adding per-trial error rates. Also, we added in our third version of 2-4-6, feedbacks of uncertainty reduction which were meant to make it easier for the participants to see that when they keep testing triples of ascending consecutive even numbers they do not reduce the uncertainty regarding the experimenter's rule. These feedbacks are similar to the results that scientists get when they do novel experiments and get closer to the truth about their subject matter of study. Testing diverse triples is akin to doing novel experiments as successful scientists do in their work. This is also the reason why we did not manipulate the origin or status of the initial 2-4-6 number triple, as was done in Van der Henst, Rossi, and Schroyens, (2002) and Caverni, Rossi, and Péris, (2005). Our manipulation was adding uncertainty reduction feedback as a stand-in of how scientists obtain novel findings when they do new experiments and how these motivate them to do further, more novel experiments. This theme of doing novel experiments to obtain more informative data that could enable scientists get closer to truth is clearly formulated in philosophy of science by Mayo's well-known error-statistical account in terms of error probabilities and severe tests (e.g. see Taper & Lele, 2004; Sarkar & Pfeifer, 2005; or Bandyopadhyay & Forster, 2011).

The notion of severe tests is central in Mayo's account according to which “Data x (produced by process G) provide a good indication or evidence for hypothesis *H* (just) to the extent that test *T* severely passes *H* with x.” (Mayo, 2005; p. 100). For a hypothesis *H* to pass a test *T* severely with data x, two things must obtain; first, data x agrees with *H*, and second, test *T* would have produced, with high probability, data that fit less well with *H* than x does, were *H* false (Mayo, 1996, 2005; Mayo & Spanos, 2011). In other words, to have reliable support for a hypothesis, the agreement between the data and the hypothesis must be difficult to obtain were the hypothesis false. One has to be sure that one has tested the different ways in which it may be wrong to infer from an agreement between the data and hypothesis that the hypothesis is true or corroborated. If we remember that the triples tested by the participants in 2-4-6 can be thought of as experiments and the experimenter's feedback can be thought of as data, it is clear that a participant who keeps testing triples of ascending even numbers fails to conduct severe tests of the hypothesis, because no matter how many triples of ascending even numbers the participant tests the experimenter feedback will always be positive yet the participant will be nowhere near discovering the experimenter’s rule. On the other hand, testing triples such as 1-3-5 or 6-4-3 are akin to conducting severe tests and hence more probably will yield useful and reliable feedback, which will steer the participant closer to discovering the rule. This is because when 1-3-5 is tested and the experimenter says it obeys the rule, the participant can infer that the rule is not about even numbers, which in effect reduces part of the uncertainty about the rule, i.e. the participant now knows that the rule they are trying to discover is not just about even numbers. Likewise, when 6-4-3 is tested and the experimenter feedback is negative, i.e. the triple does not obey the rule, the participant now knows the rule is about ascending numbers. Thus, it can be seen that when participants test diverse kinds of triples in the 2-4-6 task, this ends up reducing part of the uncertainty about the rule, which are analogous to conducting severe tests in the error-statistical sense as defined by Mayo (1996, 2005). Thus, it can be reasoned that if the participants are made aware of error probabilities and uncertainty in 2-4-6 and how these can be reduced, they may perform better at this task.

Following the reasoning above, three new versions of 2-4-6 incorporating error rates and feedback of uncertainty reduction were employed in three experiments. The manipulation of making experimenter feedback probabilistic, rather than deterministic as in the original version, has been done in a small number of previous studies by telling participants that some of the experimenter's feedback would be in error. Gorman (1989) in four experiments has told participants that between 0% and 20% of experimenter feedback would be in error, so the participants had to guess the actual probability that feedback in a given trial was in error. Participants in the control condition were not told anything about probability of erroneous experimenter feedback. So, a participant in the experimental condition who tested a given number triple could not be 100% sure whether or not the triple did indeed obey the experimenter's rule. The results of Gorman's initial three experiments showed that participants in the error rate group tested a greater number of triples and they also had the trend to replicate by testing the same triple more than once. In Gorman's experiments, a significantly greater number of participants in the control condition (with no error rates) successfully discovered the rule.

Although participants in the experiments discussed above were told that some experimenter feedback was going to be erroneous, in reality they were not given any erroneous feedback. Gorman (1989) in his fourth experiment, gave the participants the same instructions as above but this time some experimenter feedback was in fact erroneous, i.e. in some trials even though participants tested a triple that fit the rule, they were told that it did not fit the rule. The results of this experiment also showed that participants in the error group were less successful in discovering the rule, they spent more time testing triples, and tried to do more replications as compared to participants in the control condition with no error rates. According to Gorman, although these results support Lakatos's (1970) idea that better hypothesis tests can be done by trying replications, they also show that this idea is difficult to apply in more hands-on hypothesis testing situations. Another finding was that, as in a previous study by Gorman (1986), when error rates were included, participants tended to more strongly cling to their hypotheses and when they got negative feedback they assumed that the feedback was in error. Indeed, this is not very different from what researchers do in real scientific practice. Gorman has concluded from these results that, although Lakatos's falsificationist approach may be philosophically valid, it cannot explain participants' hypothesis testing strategies in problems that include error probabilities.

Another study in which error rates were incorporated in the 2-4-6 task was by Penner and Klahr (1996). Participants were told that some experimenter feedback would be subject to system error, i.e. some feedback would be negative for triples that fit the experimenter's rule and positive for some triples that do not fit the rule. As in Gorman (1989), 0% to 20% of experimenter feedback would be subject to system error. Participants in the control condition were not told anything about system error and were not given any erroneous feedback. The results showed that participants in the system error condition were significantly less successful in discovering the experimenter's rule. But these participants tested twice as many triples as those in the control condition, which replicated Gorman's (1989) findings. Another finding was that all participants in the error group found out which feedback was erroneous at least once, which showed that even though they could not discover the rule fast enough, the participants did understand the false-positive and false-negative feedback coming from the experimenter. Gorman's (1989) finding that participants tended to assume negative feedback as erroneous was also obtained to some extent by Penner and Klahr (1996). However, they suggest that how participants evaluate the negative feedback is much more complex and the available findings are not sufficient to explain how people shape their hypothesis testing strategies and how they try to find out which ones of the experimenter's feedback are erroneous in problems with possibility of error.

Although the kind of manipulation used by Gorman (1989) and Penner and Klahr (1996) does make the 2-4-6 task more similar to real contexts of hypothesis testing, it still is considerably different from actual hypothesis testing situations. A possible error rate that can be as high as 20% is not very realistic because in actual research settings a test with such a high error rate would not be taken seriously by researchers. Another issue in the above experiments is the fact that one report of error probability (varying from 0% to 20%) was given to participants for the whole experiment. Thus, the participants did not know the error probability associated with each specific trial of a generated triple but had to deal with a varying error probability associated with the totality of all their trials.

However, in real settings, researchers have to evaluate the specific error probability associated with each specific hypothesis test, the most obvious example being the use of p-values. If we think of each triple the participants test as an analog of a scientific experiment, we can see that the versions used by Gorman (1989) and Penner and Klahr (1996) are somewhat misleading for the participants and far from closely resembling real hypothesis testing situations. In a sense, the manipulation of these experiments gave participants the uncertainty of uncertainty. In contrast, the participants in our experiments were given a more direct feedback of error probability in terms of specific error rates for every trial. In experiments 1 and 2 reported here, different error rates were given to participants after each triple they tested, so they knew the specific error rate per each trial. When each trial is thought of as an experimental analogue of hypothesis testing in a real setting (science or industry), knowing the specific error rate for each trial more closely resembles the use of statistical significance values or false-positive or false-negative rates. In our third experiment, a novel manipulation was used: after each trial, the participant was given a specific percentage rate as a feedback of the extent to which they reduced the uncertainty about what the experimenter's rule may be.

Experiment 1

Methods

Sample

Turkish-speaking university students, ranging in age from 19 to 23 (mean=21.34 ; st.dev.=1.29), participated for course credit in the experiment with two groups: control group (n=38, female=26) and experimental group with error probabilities (n=34, female=24). All participants were adults and they were asked to give their consent by signing an informed consent form before the experiment began.

Procedure

The participants were run one at a time and each was first given a verbal working memory task in which the experimenter read aloud eight letters and the participants were asked to recall as many of those letters as they could. This was done to establish that participants in the groups did not differ in their normal verbal working memory capacities that could interfere with their performance in the 2-4-6 task. We used a verbal working memory task with letters instead of numbers so that the participants would not get practice with numbers before starting the 2-4-6 task. Participants were then instructed as in Wason's original paradigm described above but they were also told that experimenter feedback would be subject to error (error probability group). They were given a data form, in which there were columns from left to right, the first two to be filled with the participants' number triples and hypotheses regarding the rule, and the last two for experimenter feedback and the error rate for that trial. The data form in the control condition was the same but without the column for error rate. The participants were asked to create a new number triple and propose a hypothesis of what the rule is but they were not required to propose a new hypothesis for each trial. The experimenter took back the form, put a check mark, depending on the triple, under either "Fits the Rule" or "Does not fit the Rule" column, and wrote under the error rate column a specific rate for each trial from a list of error rates consisting of 3%, 5%, 7% or 10% following a previously set random order. For example, for a triple such as 8-10-12, the experimenter put a check mark under "Fits the Rule" column and wrote 5% under the error probability column, which meant that the probability that this feedback is in error is 5%. Although participants were told that feedback would be subject to error, in reality participants were given no erroneous feedback in experiment 1. The participants were clearly instructed that the error probability attached to the specific feedback for that trial and not to their hypothesis of what the experimenter's rule may be. The rule the participants were asked to discover was the same rule Wason (1960) used, namely "any three ascending numbers." The participants in the control group went through the identical procedure but without any instruction about or reports of error probabilities and the experimenter feedback they got about triples fitting or not fitting the rule was free of any error (control group).

Dependent Measures

In experiment 1, four major dependent measures were defined: successful rule discovery, number of triples generated by the participant, proportion of non-fitting triples (defined as the number of triples that did not fit the experimenter's rule divided by the total number of triples generated), and triple diversity. This dependent measure was defined as a measure of the diversity of the triples participants generated; if a participant who generated triples with odd numbers (1-3-5), numbers of unequal difference (4-8-10), decreasing numbers (8-6-4) etc. they got a higher triple diversity score than another participant who generated triples with only even numbers (8-10-12) or only increasing numbers (20-24-28). Triple diversity was defined as a measure of the effectiveness of a participant’s strategy in solving the problem, because given the nature of the problem and the rule they were asked to discover, the more diverse kinds of triples a participant tried the greater was the chance of discovering the rule. The number of different hypotheses regarding the rule that participants wrote down and the numbers of negative tests (falsification) and positive tests (confirmation) they did were also collected.

Results and Discussion

Since the obtained data on verbal working memory performance were not normally distributed, a Mann-Whitney U test was used, which does not assume normality. This analysis showed no significant differences between participants in the two groups with respect to verbal working memory performance (U = 588.5, p = .494). Figure 1 shows that a greater number of participants in the control group correctly guessed the rule (9 in control versus 3 in error rate group), however a Chi-Square analysis showed that this difference was not significant, χ^2(1, N=72)=2.85, p=.91.

(Insert figure 1 about here)

Mann-Whitney U tests revealed that the participants in the control condition tested a significantly greater number of hypotheses (*U*=363.0, *p*=.027). The median number of hypotheses tested in the control group was 4 compared to a median of 2 in the error probability group. Mann-Whitney U tests also revealed that there were no significant differences between the groups in the number of negative hypothesis tests (*U*=524.0, *p*=.27) or positive hypothesis tests (*U*=480.5, *p*=.521). In both groups, the participants overwhelmingly did positive tests; the percentage of negative tests in the control group was 2.57% and in the error rate group it was 1.84%. In both groups, less than 1% of the participants did a negative test before they guessed the rule. Unsurprisingly, a Mann-Whitney U test showed that correctly guessing the rule was significantly associated with testing a greater number of hypotheses; those who discovered the rule tested a median of 4 hypotheses whereas the median of hypotheses tested by those who could not discover the rule was 2.5 (*U*=186.0, *p*=.047). On the other hand, another Mann-Whitney U test revealed that doing negative tests was not significantly associated with correct rule discovery (*U*=236, *p*=.061).

The data on the dependent measures of number of triples, proportion of non-fitting triples, and triple diversity were not normally distributed, consequently Mann-Whitney U tests were conducted on all these dependent measures and effect sizes were estimated using the formula r = Z/√N (Fritz et al., 2012). The two groups did not differ with respect to the number of triples they generated, but, as seen in figure 2, the Mann-Whitney test showed that the participants in the control group had a significantly greater proportion of non-fitting triples than those in the error rate group (*U*=386.0, *p*=.002, *r*=.37). Another Mann-Whitney U test showed that correctly guessing the rule was strongly associated with generating more non-fitting triples, i.e. those who correctly guessed the rule had a greater proportion of non-fitting triples than those who did not (*U*=101.5, *p*<.001, *r*=.49).

(Insert figures 2 and 3 about here)

Taken together, these results show that participants in the control group generated a greater number of triples that did not fit the rule and this either helped them correctly guess the rule or come close to doing so. Given the rule they were trying to discover, i.e. “any three ascending numbers,” the data supported our prediction that the greater the number of different kinds of triples the participants generate the greater are their chances of discovering the rule; e.g. trying 8-10-12, then 1-3-5, then 30-55-78, etc. The dependent measure of triple diversity was defined to measure this; each participant was given a score reflecting the diversity of the triples they generated, e.g. a participant who generated triples with odd numbers, numbers of unequal difference, non-trending numbers, decreasing numbers etc., got a higher triple diversity score than a participant who generated triples with only even numbers or only increasing numbers. Indeed, a Mann-Whitney U test showed that those who correctly guessed the rule had higher triple diversity scores (*Median*=4) than those who did not (*Median* =3.00), (*U*=129.0, *p*<.001, *r*=.42). In addition, another Mann-Whitney U test showed that participants in the control group had significantly higher triple diversity scores (*Median* =4.00) than those in the error probability group (*Median* =2.5), (*U*=406.5, *p*=.006, *r*=.33). The effect sizes of the above results ranged from r=.33 to r=.49, which pointed to medium to large effects.

Overall, these findings suggest that participants in the control group performed better than those in the error probability group in generating more effective number triples and increasing their chances of discovering the rule.

Experiment 2

Methods

Sample

Turkish-speaking university students, ranging in age from 19 to 25 (mean=21.76 ; st.dev.=1.66), participated for course credit in the experiment with two groups: control group (n=38, female=26) and experimental group with actual error (n=35, female=25). All participants were adults and they were asked to give their consent by signing an informed consent form before the experiment began.

Procedure

In experiment 2, the procedure in the experimental condition was the same as the error probability group in experiment 1 but with actual error. The participants were given rule obedience feedback and specific error rates after each trial; but on certain randomly selected trials, they were given erroneous feedback, i.e. if the participant generated a triple that did not fit the rule, the experimenter said it did fit the rule and vice versa (actual error group).

Dependent Measures

The same four dependent measures from experiment 1 were analyzed; namely successful rule discovery, number of triples, proportion of non-fitting triples, and triple diversity. The performance of participants was compared to the control group with no error rates. The number of different hypotheses regarding the rule that participants wrote down and the numbers of negative tests (falsification) and positive tests (confirmation) that they did were also collected.

Results and Discussion

There were no significant differences between participants in the two groups with respect to verbal working memory performance as shown by a Mann-Whitney U test ($U = 611, p = .531$).

A Chi-Square test showed that significantly fewer participants in the actual-error group correctly guessed the rule compared to the control group (9 versus 2), χ^2(1, N=73)=4.59, *p*=.032). As in experiment 1, Mann-Whitney U tests in experiment 2 revealed that the participants in the control condition tested a significantly greater number of hypotheses (*U*=387.5, *p*=.017); the median number of hypotheses tested in the control group was 4 compared to a median of 2 in the actual error group. Mann-Whitney U tests also revealed that there were no significant differences between the groups in the number of negative hypothesis tests (*U*=567.5, *p*=.826) or positive hypothesis tests (*U*=496.0, *p*=.315). As in experiment 1, the participants in both groups in experiment 2 overwhelmingly did positive tests; the percentage of negative tests in the control group was 2.57% and in the actual error group it was 1.04% and in both groups, less than 1% of the participants did a negative test before they guessed the rule. Also as in experiment 1, correctly guessing the rule was significantly associated with testing a greater number of hypotheses; those who discovered the rule tested a median of 4 hypotheses whereas the median of hypotheses tested by those who could not discover the rule was 2 (*U*=175.0, *p*=.042).

A Mann-Whitney U test also showed that the participants in the control group had a significantly greater proportion of non-fitting triples than those in the actual error group (*U*=432.5, *p*=.007, *r*=.45). Another Mann-Whitney U test showed that correctly guessing the rule was strongly associated with generating more non-fitting triples, i.e. those who correctly guessed the rule had a greater proportion of non-fitting triples than those who did not (*U*=120.5, *p*<.001, *r*=.60). The participants in these groups did not differ in the number of triples generated but the participants in the control group achieved significantly higher triple diversity scores (*Median* =4) than those in the actual error group (*Median* =2), (*U*=424.5, *p*=.007, *r*=.32). In addition, a Mann-

Whitney U test showed that those who correctly guessed the rule had higher triple diversity scores (*Median*=4) than those who did not (*Median* =3.00), (*U*=142.5, *p*=.002, *r*=.52).

(Insert figures 4, 5, and 6 about here)

Taken together, the results of experiments 1 and 2 suggest that performance on the 2-4-6 task is impaired when experimenter feedback is framed to include probabilistic error, with or without actual error, but arguably more so when there is actual error. These results partially replicate the general findings of Gorman (1989) and Penner and Klahr (1996), which suggest that when there is a possibility of error in feedback, the participants do not change their working hypothesis about the rule when they get a negative feedback and they do not try new kinds of triples. As a result, their chances of successful rule discovery decreases. However, Gorman (1989) and Penner and Klahr's (1996) finding that participants in the error probability group tried more triples was not replicated here. In both experiments 1 and 2, there were no significant differences between the groups in the number of triples generated. The reason for this may have been that they were given an error rate for each trial, which could have discouraged them to try more triples and/or it may have been due to a cultural difference. Doing experiments 1 and 2 with non-Turkish speaking participants would in fact help further study this possibility of cultural differences affecting performance on the 2-4-6 task.

Experiment 3

Inspired by Mayo's error-statistical account (Mayo, 1996; 2005) as well as picking up on cues from the above findings, in experiment 3, we devised a version of the 2-4-6 task to include a kind of experimenter feedback that would steer participants to generate a greater number of different kinds of triples.

Methods

Sample

Turkish-speaking university students, ranging in age from 19 to 33 (mean=21.62 ; st.dev.=2.53) participated for course credit in the experiment with two groups: control group (n=38, female=26) and experimental group with uncertainty reduction feedback (n=32, female=21). All participants were adults and they were asked to give their consent by signing an informed consent form before the experiment began.

Procedure

The procedure in the experimental group was the same as the error rate group in experiment 1, but instead of error rates, the participants were given uncertainty reduction feedback for each triple they generated. Before they started, the participants were clearly instructed that there was an uncertainty about what the experimenter's rule is but that they could reduce that uncertainty with the triples they generate. For each triple, the experimenter gave the standard rule-obedience feedback, but in addition, uncertainty reduction feedback was given for each triple in terms of a percentage. For example, if the participant generated a triple of consecutively ascending even numbers (e.g. 8-10-12) the uncertainty reduction was 0% but if the triple was all odd numbers (e.g. 3-5-7), or of numbers of unequal difference (e.g. 4-9-12), the uncertainty reduction was 3%. If the participants generated triples of the same number (e.g. 4-4-4), or decreasing (e.g. 14-12-10) or non-trending numbers (e.g. 2-5-3), uncertainty reduction feedback was 5%. The feedbacks of uncertainty reduction were not additive, but when the participant's triple included two of these aspects, e.g. 3-7-15, odd numbers as well as numbers of unequal difference, the uncertainty reduction feedback was increased to 5%. Of course, these reports were not meant as genuine reduction in uncertainty; the participants were not given any instructions about any upper limit on the numbers they could use so they could use all numbers in their triples. Also, the rule they tried to discover could have been about all numbers. As Cantor's arithmetic of cardinality dictates, infinity minus infinity equals infinity, in strictly mathematical terms uncertainty cannot be reduced by testing different kinds of triples. Nonetheless, testing different kinds of triples does increase the chances of discovering the rule, this is why participants were given these informal reports of uncertainty reduction to reward any tendency to test diverse number triples.

Dependent Measures

The four major dependent measures from experiment 1 were analyzed, namely successful rule discovery, number of triplets generated by the participant, proportion of non-fitting triples, and triple diversity. The number of different hypotheses regarding the rule that participants wrote down and the numbers of negative tests (falsification) and positive tests (confirmation) they did were also collected.

Results and Discussion

A Mann-Whitney U test showed no significant differences between participants in the two groups with respect to verbal working memory performance (*U = 563.5, p = .577*). A Chi-Square analysis showed that significantly fewer participants in the uncertainty-reduction group correctly guessed the rule, compared to the control group (9 versus 2), χ^2(1, N=70)=3.98, *p*=.046).

(Insert figure 7 about here)

In contrast to experiments 1 and 2, Mann-Whitney U tests in experiment 3 revealed that there were no significant differences between the groups in the number of hypotheses tested

(*U*=521.0, *p*=.926), negative hypothesis tests (*U*=497.0, *p*=.517) or positive hypothesis tests (*U*=481.5, *p*=.540). But, as in experiments 1 and 2, the participants in both groups in experiment 3 overwhelmingly did positive tests; the percentage of negative tests in the control group was 2.57% and in the error rate group it was 2.33% and in both groups, less than 1% of the participants did a negative test before they guessed the rule. Also, Mann-Whitney U tests revealed no significant differences between those who correctly discovered the rule and those that did not in the number of hypotheses they tested (*U*=193.0, *p*=.129), or the number of negative tests (*U*=223.0, *p*=.132) or the number positive tests they did (*U*=265.0, *p*=.855).

In contrast to the findings from experiments 1 and 2, there were no significant differences between the control group and the uncertainty-reduction group in the number of triples generated (*U* = 606, *p* = .981), proportion of non-fitting triples (*U* = 605.5, *p* = .976), and scores of triple diversity (*U* = 604, *p* = .961).

Correctly guessing the rule was strongly associated with trying a greater number non-fitting triples (*U* = 148, *p* < .005, *r* = .35) and also with higher triple diversity scores (*U* = 172.5, *p* < .02, *r* = .30). As these are measures associated with success in the 2-4-6 task, these findings suggest that when experimenter feedback in 2-4-6 is framed to include feedbacks of uncertainty reduction on the basis of the diversity of triples generated, the participants' performance is not as impaired as when the experimenter feedback on rule obedience of the triples is subject to probabilistic error. Aside from correct guesses of the rule, the findings that those in the uncertainty-reduction group generated as many non-fitting triples, and achieved as high similar triple diversity scores as the control group show that the adverse effects of error rates were not observed when uncertainty reduction feedback for each triple is added to rule obedience feedback.

(Insert figures 8 and 9 about here)

General Discussion

The impaired performance in experiments 1 and 2 in discovering the rule when error rates were added to the 2-4-6 task may be seen as partial replications of the findings of Gorman (1989) and Penner and Klahr (1996). However, the findings of experiment 3 suggest that when uncertainty reduction feedback is added to the 2-4-6 task, participants' performance in generating non-fitting triples and triple diversity scores are not as impaired compared to experiments 1 and 2 with error rates. Although participants in the uncertainty-reduction group were significantly less successful in discovering the rule, the fact that they generated as many non-fitting triples and achieved similar triple diversity scores as the control group suggests that the impairing effects of error rates are not observed with uncertainty reduction feedback. In addition, as can be seen in figures 2 and 5, participants in the error rate groups in experiments 1 and 2, respectively, generated significantly smaller numbers of non-fitting triples compared to those in the control group without error rates. In contrast, as can be seen in figure 8, participants in the uncertainty reduction group in experiment 3, generated as many non-fitting triples as those in the control group. Parallel comparisons can be made regarding triple diversity scores across three experiments. Namely, figures 3 and 6, respectively show that in experiments 1 and 2, participants in the error rate groups achieved significantly lower triple diversity scores than those in the control group. In contrast, figure 9 shows that participants in the uncertainty reduction group in experiment 3, achieved similar triple diversity scores as those in the control group. When put together, these findings show that uncertainty reduction feedback does not impair performance in the 2-4-6 task in the way error rates do as clearly seen in experiments 1 and 2. In all three experiments, as shown in figures 2, 3, 5, 6, 8, and 9, generating a greater number of non-fitting triples and achieving higher triple diversity scores are significantly associated with discovering the rule. These are the same factors that were negatively affected by error rates while they were not impaired when participants were provided with uncertainty reduction feedback.

This result in turn lends support to the possibility that introducing the notion of uncertainty reduction may improve the kinds of reasoning and strategies associated with success in rule discovery tasks. Also, this point can be made clearly when one looks at the 2-4-6 task from the perspective of dual process models of reasoning (e.g. Evans, 2003; Kahneman, 2011) in which system 1 leads the participant to give the automatic and spontaneous but non-effective response to a task whereas system 2 gives the more controlled, analytically sound yet slower response as they try to solve a problem. It may be that those participants who are overly primed by the initial triple 2-4-6 and generate mostly similar triples of even numbers ascending by 2 are operating in a mode compatible with system 1 whereas the participants who generate different types of triples and hence achieve higher triple diversity scores may be closer to what the more analytic system 2 would do. We suggest that adding uncertainty reduction feedback to the 2-4-6 task may trigger the greater involvement of system 2 in this task and thus lead to more effective triple tests and improve the chances of success. This can also be related to the influential analyses of the 2-4-6 task offered by Klayman and Ha (1987); what appears to be important and conducive to success in this task is not whether or not one employs a falsificationist or confirmationist strategy but rather whether or not one generates a sufficiently diverse set of triples to receive expected or unexpected negative feedback regarding the rule. This would be be true regardless of whether one construes the 2-4-6 task as an inductive or a deductive task. The results of the experiments reported here show that uncertainty reduction feedback does not impair the generation of diverse sets of triples as error rates do. In further experiments, more salient versions of uncertainty reduction feedback may motivate the testing of diverse triples the results of which would trigger the involvement of system 2 and thus improve the chances of success in rule discovery tasks.

Let us also remember that doing more effective tests in the 2-4-6 task are akin to conducting severe tests as described in Mayo’s error-statistical account. Thus, greater numbers of non-fitting triples and higher triple diversity scores in the 2-4-6 task, when it is expanded to include uncertainty reduction feedback, may be construed as behavioral analogs of conducting a severe tests. Just like severe tests provide scientists with more informative data, generating a greater number of non-fitting triples and achieving higher triple diversity scores provide the participants with more informative feedback. As such, these findings may build bridges between a contemporary and well-known account of inference in philosophy of science and the cognitive psychology of scientific reasoning. When we construe experimentation in science as a practice of finding ways to reduce uncertainty in contexts of testing and discovery, then the findings from these experiments can potentially be instrumental in better understanding actual inference and decision settings in science or industry. Defining uncertainty reduction as a cognitive function that is essential for success in rule discovery tasks, as well as other tasks involving uncertainty, may be fruitful in achieving a richer and more naturalistic account of the kinds of reasoning involved in contexts of discovery and inference.

References

Bandyopadhyay, P.S & Forster, M.R. (2011). *Handbook of Philosophy of Science Volume 7: Philosophy of Statistics*. Elsevier.

Carruthers, P., Stich, S., Siegal, M. (2002). *The Cognitive Basis of Science*. Cambridge: Cambridge University Press.

Caverni, J.P., Rossi, S., & Péris, J.L. (2005). “How to defocus in hypothesis testing: Manipulating the status of the initial triple in the 2-4-6 problem.” In V. Girotto & P.N. Johnson-Laird (Eds.), *The shape of reason*. Hove: Psychology Press.

Evans, J. St. B. T. (2003). "In two minds: dual-process accounts of reasoning." *Trends in Cognitive Sciences*, *7* (10): 454–459.

Evans, J. St. B. T. (2014). “Reasoning, biases and dual processes: The lasting impact of Wason (1960).” *The Quarterly Journal of Experimental Psychology*, 67, 1–17.

Feist, G.J. (2006). *The Psychology of Science and the Origins of the Scientific Mind*. New Haven: Yale University Press.

Fritz, C.O., Morris, P.E., & Richler, J.J. (2012). "Effect size estimates: Current use, calculations, and interpretation." *Journal of Experimental Psychology: General*, 141(1), 2-18.

Gorman, M.E. (1986). “How the possibility of error affects falsification on a task that models scientific problem solving.” *British Journal of Psychology*, 77, 85-96.

Gorman, M.E. (1989). “Error, falsification, and scientific inference: An experimental investigation.” *The Quarterly Journal of Experimental Psychology*, 41A, 385–412.

Holyoak, K.J. & Morrison, R. G. (2012). *The Oxford Handbook of Thinking and Reasoning*. New York: Oxford University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Klayman, J. & Ha, Y.-w. (1987). “Confirmation, disconfirmation, and information in hypothesis testing.” *Psychological Review*, 94(2), 211–228.

Lakatos, I. (1970). "Falsification and the methodology of scientific research programs." Lakatos, I. ve A. Musgrave (Eds.) *Criticism and the Growth of Knowledge*, 91-196. New York: Cambridge University Press.

Mayo, D. (1996). *Error and the Growth of Experimental Knowledge*. Chicago, IL: The University of Chicago Press.

Mayo, D. (2005). "Evidence as Passing Severe Tests: Highly Probable versus Highly Probed Hypotheses." in Achinstein, P. (ed.) *Scientific Evidence: Philosophical Theories & Applications*, pp. 95 – 127. Baltimore, MD: The Johns Hopkins University Press.

Mayo, D. & Spanos, A. (2011). "Error Statistics." In Prasanta S. Bandyopadhyay and Malcolm Forster (eds.) *The Handbook of Philosophy of Science, Volume7: Philosophy of Statistics*. Amsterdam, The Netherlands: Elsevier Publishers.

Penner, D.E. & Klahr, D. (1996). "When to trust the data: Further investigations of system error in a scientific reasoning task." *Memory and Cognition*, 24, 655-668.

Sarkar, S. & Pfeifer, J. (2005). *Philosophy of Science: An Encyclopedia*. London, UK: Routledge.

Sternberg, R.J. & Ben-Zeev, T. (2001). *Complex Cognition: The Psychology of Human Thought*. New York: Oxford University Press.

Taper, M. & Lele, S. (2004). *The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations*. Chicago, IL: University of Chicago Press

Van der Henst, J.B., Rossi, S., & Schroyens, W. (2002). "When participants are not misled they are not so bad after all: A pragmatic analysis of a rule discovery task." In W.D. Gray & C. Shunn (Eds.) *Proceedings of the 24th Annual Conference of the Cognitive Science Society* (pp. 902-907). Mahwah, NJ: Laurence Erlbaum Associates.

Wason, P. C. (1960). "On the failure to eliminate hypotheses in a conceptual task." *Quarterly Journal of Experimental Psychology*, 12, 129–140.

Appendix: Figures

Figure 1: Rule discovery performance (task success) in Experiment 1. “Count” on the y-axis refers to number of participants.

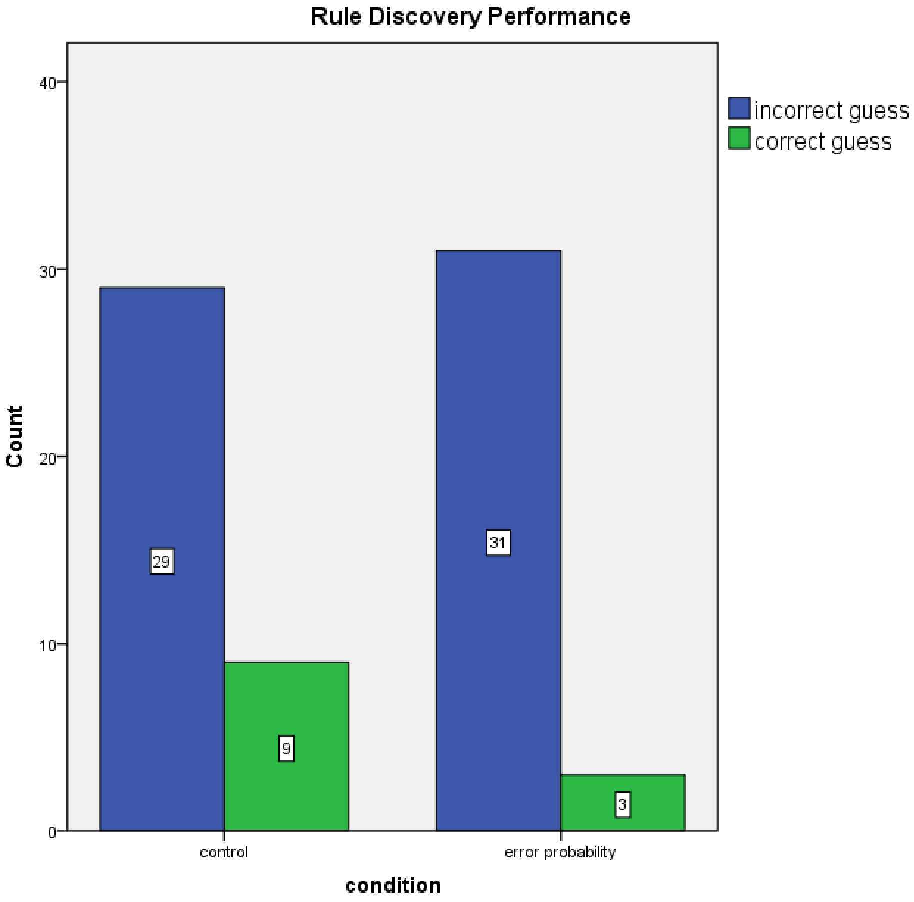


Figure 2: Proportion of non-fitting number triples generated by participants in Experiment 1. The y-axis represents the mean rank values as computed in the Mann-Whitney U analyses, where higher values refer to greater quantities.

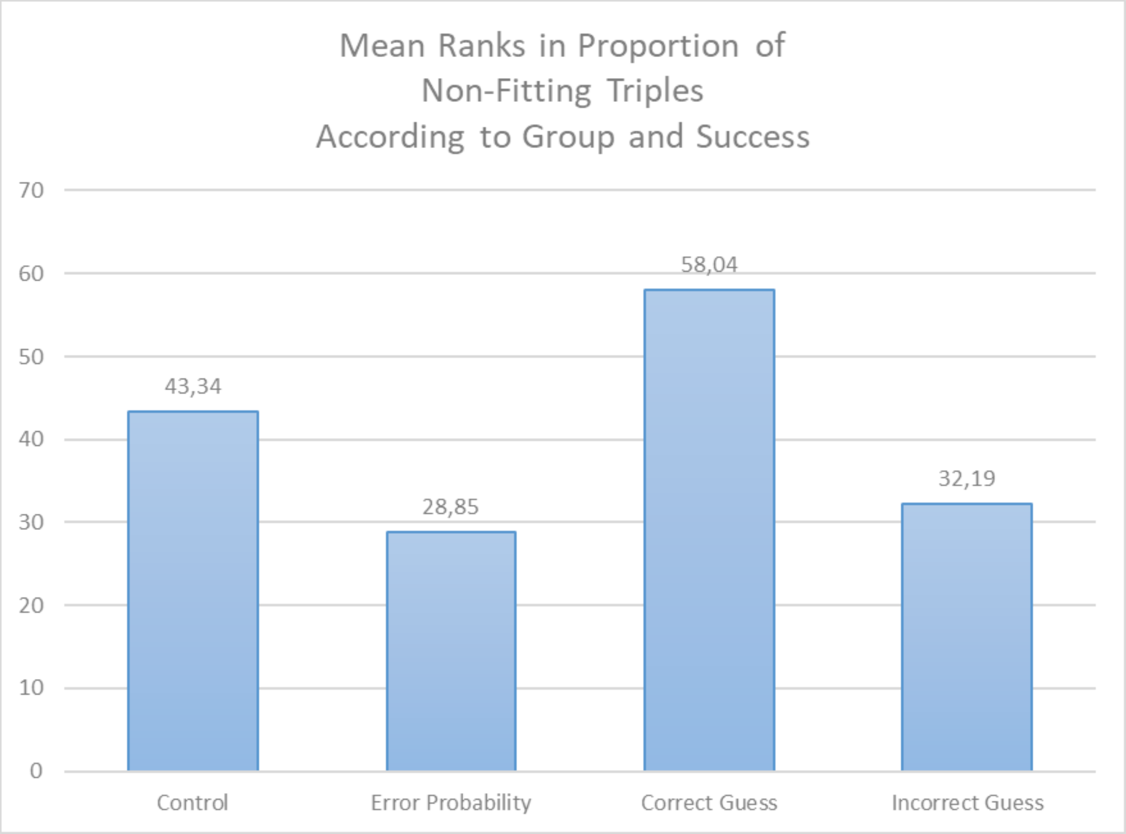


Figure 3: Triple diversity scores according to group and task success in Experiment 1. The y-axis represents the mean rank values as computed in the Mann-Whitney U analyses, where higher values refer to greater quantities.

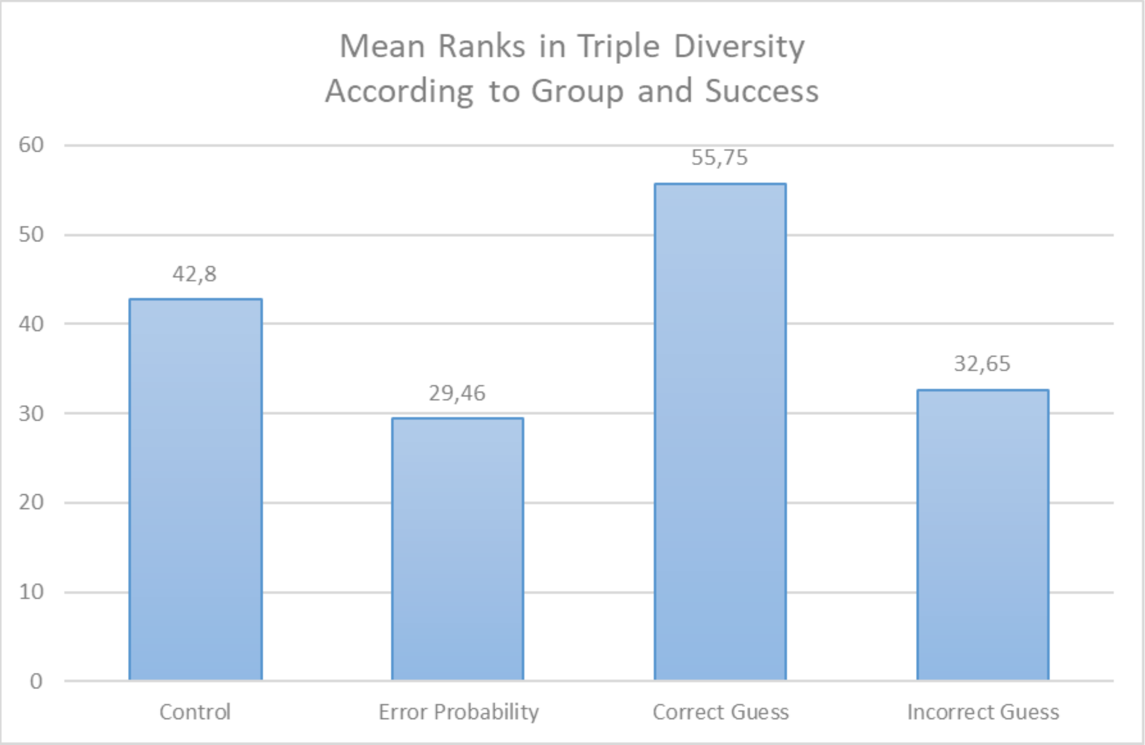


Figure 4: Rule discovery performance (task success) in Experiment 2. “Count” on the y-axis refers to number of participants.

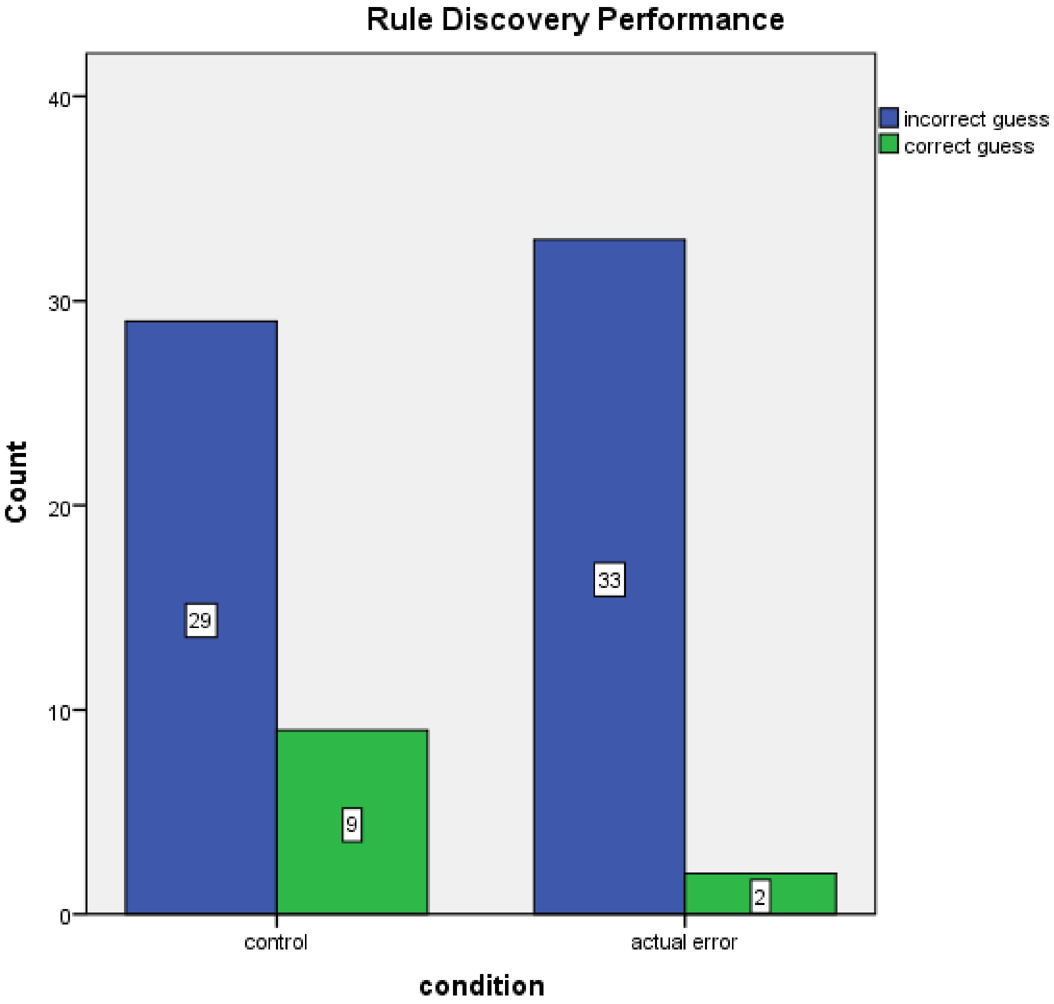


Figure 5: Proportion of non-fitting triples according to group and task success in Experiment 2. The y-axis represents the mean rank values as computed in the Mann-Whitney U analyses, where higher values refer to greater quantities.

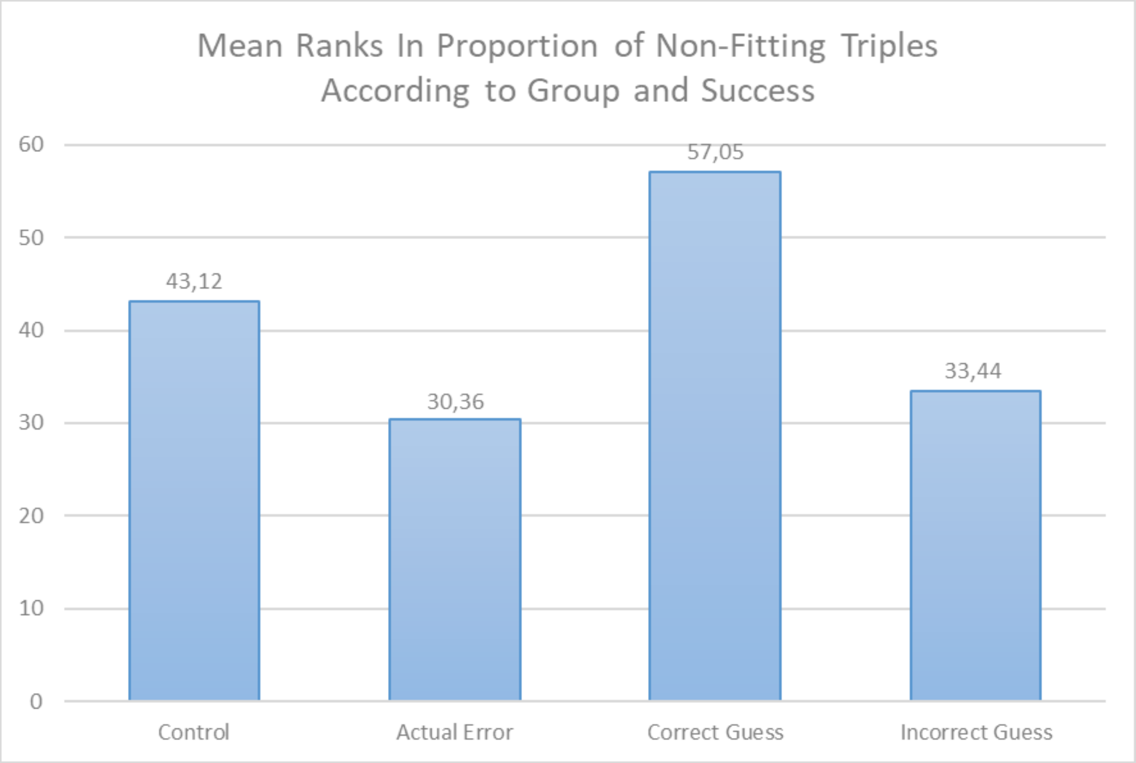


Figure 6: Triple diversity scores according to group and task success in Experiment 2. The y-axis represents the mean rank values as computed in the Mann-Whitney U analyses, where higher values refer to greater quantities.

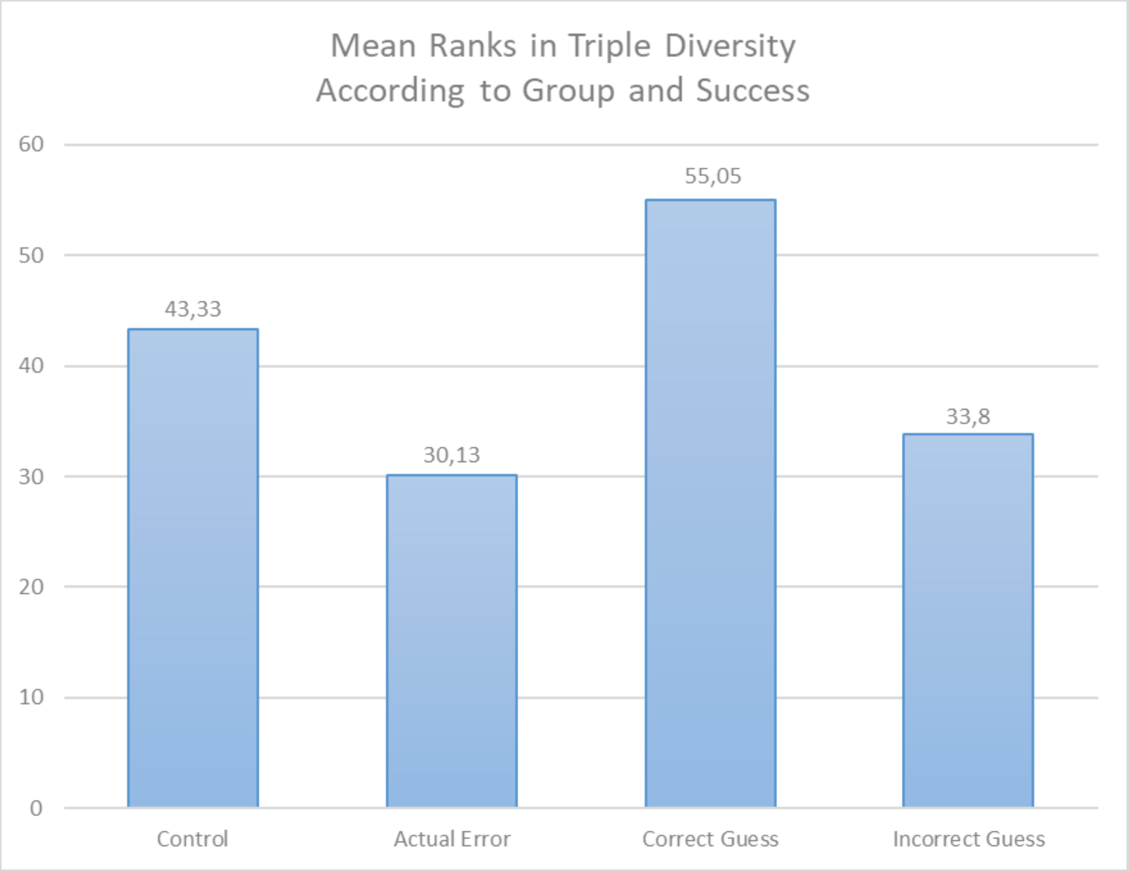


Figure 7: Rule discovery performance (task success) in Experiment 3. “Count” on the y-axis refers to number of participants.

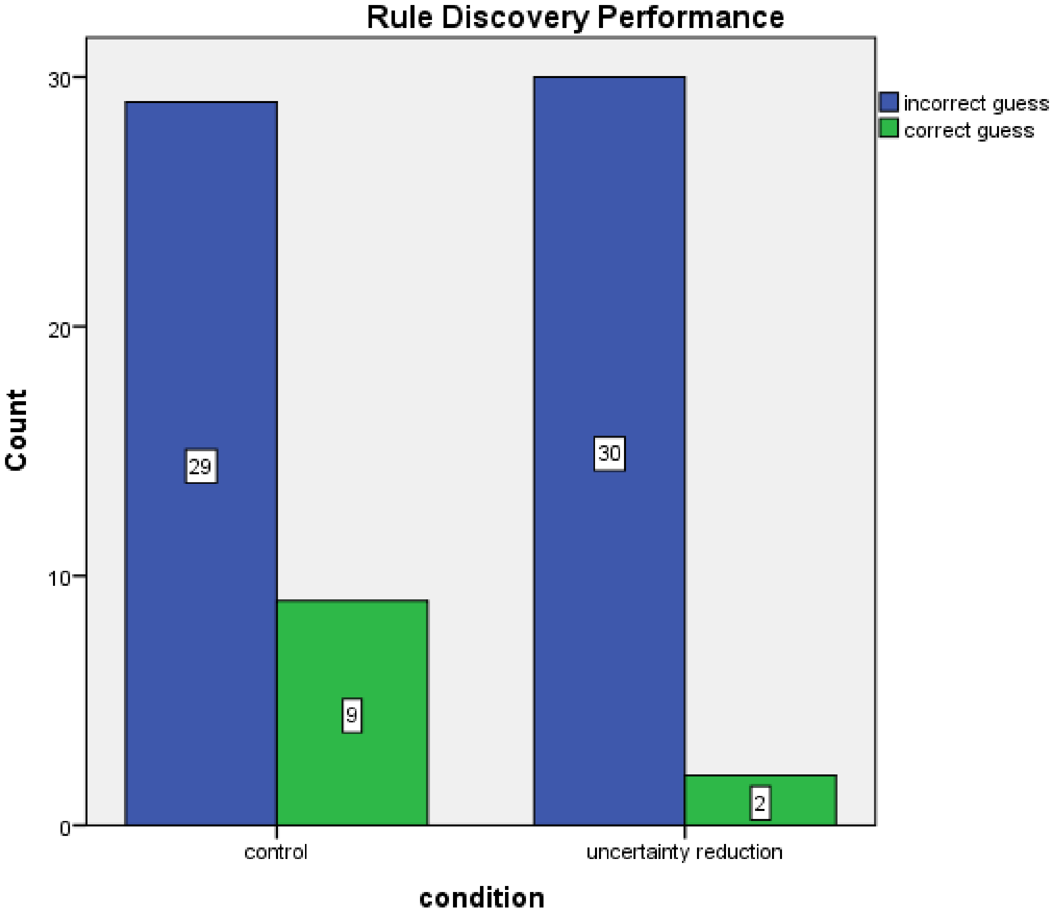


Figure 8: Proportion of non-fitting number triples generated by participants in Experiment 3. The y-axis represents the mean rank values as computed in the Mann-Whitney U analyses, where higher values refer to greater quantities.

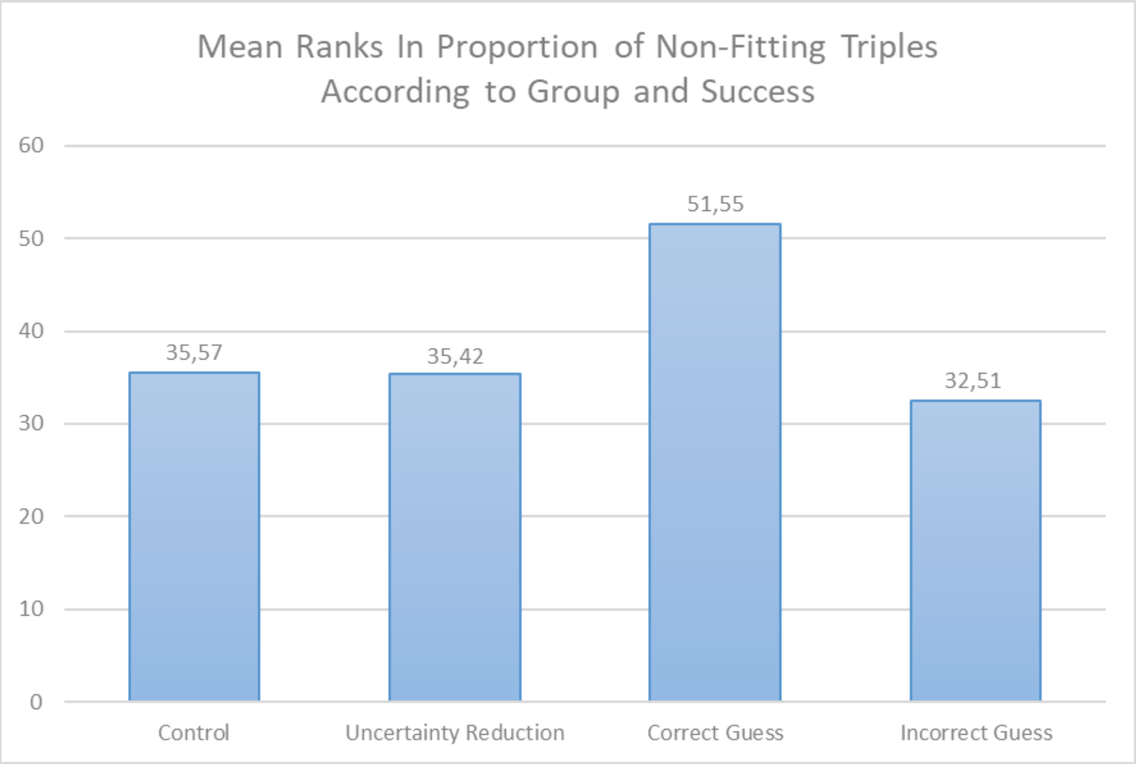
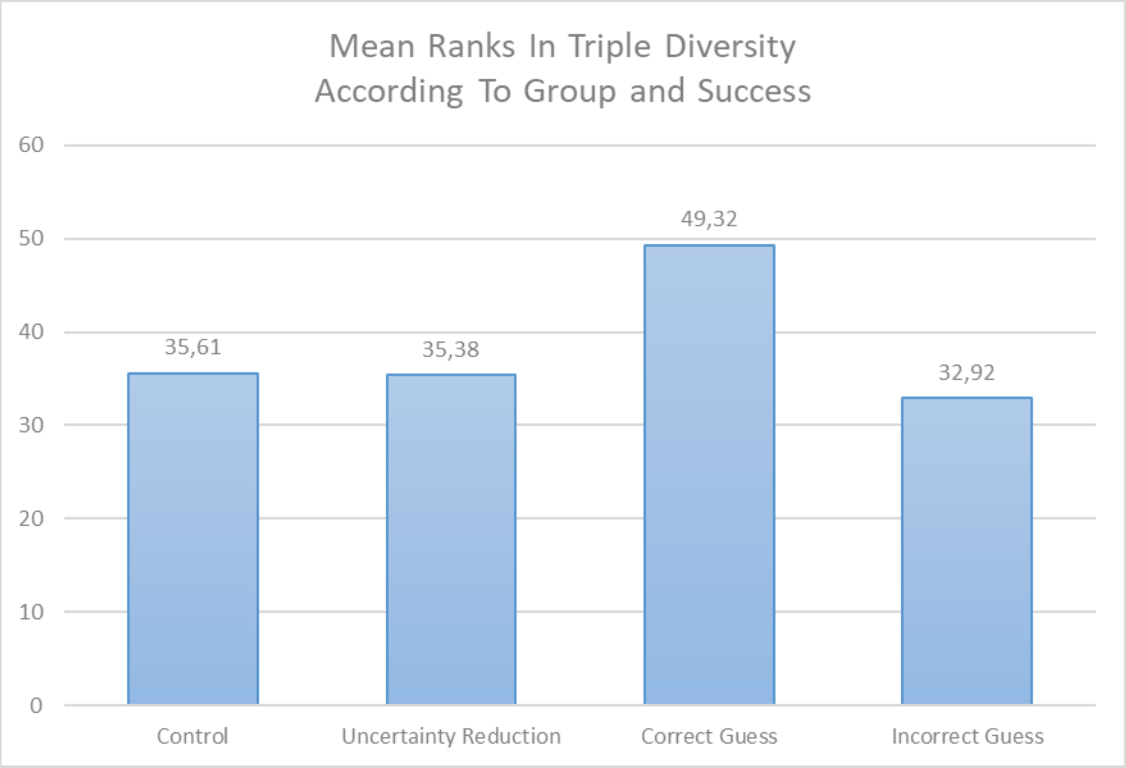


Figure 9: Triple diversity scores according to group and task success in Experiment 3. The y-axis represents the mean rank values as computed in the Mann-Whitney U analyses, where higher values refer to greater quantities.


Mean Ranks In Triple Diversity
According To Group and Success
Control
Uncertainty Reduction
Correct Guess
Incorrect Guess