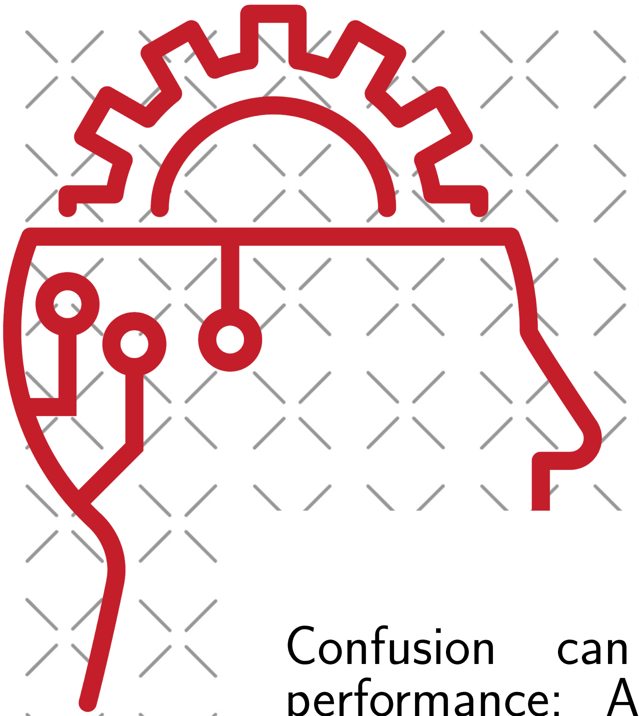
NIM Nuremberg Institute for Market Decisions



Confusion can improve cognitive performance: An experimental study using automatic facial expression analysis

Lotta Blum, Anja Dieckmann, Matthias Unfried

NIM WORKING PAPER SERIES, No. 8 / 2020

Publisher: Nürnberg Institut für Marktentscheidungen e.V. / *Founder and Anchor Shareholder of GfK SE*

Confusion can improve cognitive performance: An experimental study using automatic facial expression analysis

Lotta Blum[†] Anja Dieckmann[*†] Matthias Unfried[†]

Abstract— *In educational psychology, previous research has shown a positive association between confusion and cognitive performance, which has been attributed to "deeper learning" processes triggered by the experience of confusion. In the present study we explore the nature of this association by testing whether there is a causal impact of confusion on cognitive performance. We experimentally induced confusion in two different types of tasks – a memory task and an attention task – by incongruent information and discrepancies. Using affective computing software, we measured the facial reaction to the experimental manipulation in form of Action Unit 4 (AU4) as main indicator for confusion. Subsequently, we analyzed the impact of confusion on contextual performance, that is, performance in the same task following the experimental manipulation, as well as on general cognitive performance, that is, performance in the Cognitive Reflection Test (Frederick, 2005). The results indicate that confusion leads to activation of AU4, and that confusion positively impacts performance in the memory task as well as general cognitive performance, but not performance in the attention task. The findings suggest that confusion may lead to deeper and more reflective information processing, which can inform the design of trainings in education settings, or even decision support tools in work settings aimed at disrupting mental routines and thus activating analytic thinking.*

Keywords— *Confusion; Cognitive Performance; Affective Computing; Facial Expressions; Decision Making*

1 Introduction

> A man may be absorbed in the deepest thought, and his brow will remain smooth until he encounters some obstacle in his train of reasoning, or is interrupted by some disturbance, and then a frown passes like a shadow over his brow. (Darwin, 1872, p. 223)

Almost a century and a half ago, Charles Darwin described the facial expression of confusion in his book *The Expression of Emotions in Man and Animals.* He observed that incongruencies or disruptions in thinking frequently come along with frowning. According to Piaget (1952), these incongruencies are often accompanied by feelings of cognitive disequilibrium and lead to an affective reaction. Since then, many authors have elaborated the concept of cognitive disequilibrium (Berlyne, 1960; Chinn and Brewer, 1993; Collins et al., 1975; Festinger, 1957; Graesser and Olde, 2003; Laird et al., 1987; Mandler, 1975). Less investigated are the affective states that accompany a cognitive disequilibrium and their impact on thinking processes. The present study focuses on one of these affective states, namely confusion, which is associated with thinking processes and outcomes. Commonly, confusion reflects a loss of understanding and is associated with mistakes (Durso and Gronlund, 1999). This conception may be one-sided, as the impact of confusion on information processing is complex, and can sometimes also be beneficial. We focus on the mechanism of confusion and investigate if it can enhance performance and lead to deeper processing. The present study addresses these questions by inducing confusion experimentally in two different tasks and measuring it with both automatic facial analysis and self-report. Furthermore, two different performance measures (task-specific and general) were collected.

[*] Corresponding author, contact: anja.dieckmann@nim.org

[†] Nürnberg Institut für Marktentscheidungen e.V., Behavioral Science Research Group, Nuremberg, Germany

Citation: Lotta Blum, Anja Dieckmann and Matthias Unfried (2020). Confusion can improve cognitive performance: An experimental study using automatic facial expression analysis. NIM Working Paper Series, No. 8 / 2020.

1.1 Confusion and its facial expression

Confusion occurs when new and unexpected events interrupt ongoing cognitive activity and the interruptions cannot be resolved immediately or be integrated into existing mental models (Mandler, 1975, 1984, 1990). According to Brosch and Scherer's (2009) Component Process Model, confusion is triggered by the appraisals of high novelty and low coping potential.

The facial expression of confusion is easy to recognize for observers. Charles Bell (as cited in Darwin, 1872) classified it as the most notable emotion expression of the human face. In their observational study about facial expressions in everyday contexts, Rozin and Cohen (2003) found that confusion was one of the most prevalent expressions that primarily involved the eyebrow region. A more systemic approach was used by Craig et al. (2008). They applied the *Facial Action Coding System* (FACS; Ekman and Friesen, 1978) to detect confusion. FACS is an exhaustive system for describing facial movement by assigning 44 different Action Units (AUs) to every muscle movement the human face is capable of. The most frequently observed AU during states of confusion was AU4, the lowering of the eyebrows, which is consistent with Darwin's description as well as Rozin and Cohen's findings.

In the last two decades, emotion detection increasingly became a subject of research in affective computing, which is the development of computer systems and software to detect, interpret, and simulate human affect or emotions. One big strand of this literature is on the inference of emotions and emotional components (e.g., appraisals) from facial expressions.

To detect AU4, a software developed by the Nuremberg Institute for Market Decisions (NIM), the Centre International des Sciences Affective (CISA) at the University of Geneva, and the Fraunhofer Institute for Integrated Circuits (Fraunhofer IIS) was used (cf. Seuss et al., 2020). It automatically infers 22 different AUs from webcam recordings. For the detection of AUs the software extracts two different types of information, namely geometric and texture-based measurements, to estimate the intensities of the AUs. While geometric features are derived from the positions of prominent facial feature points (e.g., the inner endpoints of the eyebrows, which for AU4 typically move closer and down), the texture-based measurements track changes in shadows produced by expression lines (e.g., vertical lines in the center of the forehead in case of AU4; see Figure 1).

Seuss et al. (2020) report that the point-biserial correlation between AU4 presence or absence coded manually by certified FACS coders and the automatic detection of AU4 intensity by the proprietary software is r = .43. The authors consider this result as good, albeit far from perfect. Hence, an additional aim of this study is to examine if the detection software is sensitive enough to capture signs of confusion by detecting increased levels of AU4 in respondents during exposure to confusing stimuli. Furthermore, we additionally included post-hoc ratings of experienced confusion.

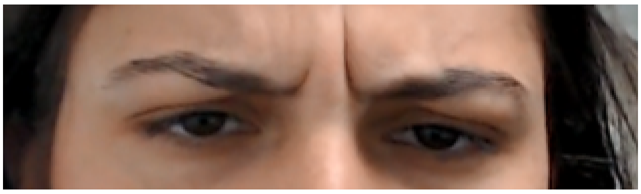


Figure 1: Typical expression of AU4.

1.2 Impact on performance

Because of the high prevalence of confusion in learning contexts (Lehman et al., 2008; D'Mello et al., 2010; D'Mello, 2013), several studies analyze the relation between confusion and learning. While some authors found negative correlations (Rodrigo et al., 2009; Schneider et al., 2015; Richey et al., 2019), other studies found a positive relationship between confusion and learning. D'Mello and Graesser (2011a) argue that confusion may increase attention and motivate effort to "deliberate, problem solve, and restructure" (p. 303) and thus lead to "deeper learning" (Craig et al., 2004). For example, Craig et al. (2004) conducted an observational online study in which they FACS-coded facial expressions to infer different affective states of learners and measured the performance using pre- and posttest. They indeed showed that confusion correlates positively with learning gains. Similar results are reported in other studies (Graesser et al., 2007; D'Mello and Graesser, 2011b).

According to this branch of research, however, confusion is not expected to be always beneficial for learning and cognitive performance. D'Mello and Graesser (2012), argue that confusion has to be contextually connected with the learning activity, and VanLehn et al. (2003) point out that confusion has to be ultimately resolved in order to improve the performance. Thus, there is some empirical support for the hypothesis that when contextually coupled with learning, confusion plays an important and mostly positive role during complex thinking processes.

However, the majority of studies focuses merely on the correlation between confusion and performance but did not investigate the causal role of confusion (Craig et al., 2004; D'Mello and Graesser, 2011b; Forbes-Riley and Litman, 2009, 2010, 2011; Graesser et al., 2007; VanLehn et al., 2003). For example, more engaged and eager learners might simply express confusion more frequently, which could account for the reported positive association. Moreover, it remains unclear whether increased unspecific attention and effort alone exert a positive influence on task performance, or whether indeed deeper learning takes place and drives the reported beneficial effects of confusion. Finally, there are also inconsistent results showing a negative relationship. This underlines the importance of further research focusing on the causal relationship between confusion and cognitive performance. The present study tries to fill this gap in the literature. In particular, the study focuses on the question if experimentally induced confusion actually increases contextual performance. Two tasks are selected to address this question, one requiring comprehension and memorization of written information (i.e., a typical learning task), which may profit from deeper learning, and another requiring attention and effort (i.e., detecting and counting visual target stimuli among distractors), where deep learning is not required.

Furthermore, the impact of confusion on general (i.e., context-unrelated) performance is analyzed. For a possible influence of confusion on general cognitive performance, we refer to Kahneman (2011). Kahneman postulated two coexisting modes of thinking that are involved in information processing. System 1 is associative and fast, working constantly but unconsciously, without any effort or volitional control, while System 2 draws attention to demanding cognitive activities and works slowly, analytically, and reflectively. System 1 gets activated as an immediate reaction to a stimulus. When an event is registered that contradicts the expectations of System 1, System 2 is activated (Kahneman, 2011). For instance, when information is difficult to process, the experience of difficulty or disfluency in reasoning can serve as a metacognitive cue that the intuitive response is possibly wrong and prompt deeper processing by System 2 (Alter et al., 2007). At least for non-routine tasks, System 2 processing can be beneficial. Once triggered, System 2 processing may lead to an increase in general cognitive performance, independent of the concrete task.

It should be noted that, although the authors of the before-mentioned studies in learning and educational psychology stress the importance of contextual connection between confusion and the learning task, there is some similarity between deep learning and System 2 processes. Nevertheless, to the knowledge of the authors, there is no contribution in the literature that addresses the impact of confusion on general, task-unrelated cognitive performance, implicating the activation of analytical System 2 processes. By taking a closer look at a potential general effect of confusion on performance, the findings could be used to systematically activate analytic reasoning strategies and thus help to avoid possible cognitive biases in decision making resulting from fast mental routines that otherwise would remain unquestioned.

1.3 Research objective and hypotheses

The purpose of this study is to analyze the impact of confusion on both contextual and general cognitive performance. Confusion is elicited by an unexpected event that cannot be resolved immediately but is resolved before the cognitive performance is measured. The experimentally triggered confusion is expected to lead to a facial reaction, that is, activation of AU4, which is analyzed using the above-mentioned software. As a side question, this analysis will show if the software proves sensitive enough to capture the confusion-induced changes in AU4.

To answer the research question we follow the strand of literature which showed a positive correlation between confusion and performance. As to the knowledge of the authors the causal relation remains largely unexplored, the present study addresses this gap in the literature by experimentally manipulating confusion, and thus is able to causally link confusion and performance. In particular, we expect confusion to increase performance in a memory task by fostering deep learning, and we expect confusion to increase performance in a task that requires diligence and effort by heightening attention. Additionally, this study focuses not only on context-related cognitive performance in two different cognitive tasks, but also on general, task-unrelated performance, which leads to the following three hypotheses:

Hypothesis 1 *Confusion has a positive impact on contextual performance in a memory task.*

Hypothesis 2 *Confusion has a positive impact on contextual performance in an attention task.*

Hypothesis 3 *Confusion has a positive impact on general, not task-related cognitive performance.*

2 Method

2.1 Participants

72 participants (33 female) completed the experiment. Participants' age ranged from 19 to 31 years (M = 23.8, SD = 2.48). 57 % of the respondents were students, 10 % were apprentices, 22 % employees, 3 % pupils and 8 % declared another occupation. Participants were recruited via social media.

2.2 Design and Procedure

The fully computer-based experiment was implemented using *oTree* (Chen et al., 2016), an open-source platform for decision experiments. It was conducted in a laboratory with six booths. Participants were randomly assigned session-wise to experimental (EG) or control group (CG) and were asked to solve two different types of tasks, A and B. In the EG confusion was evoked in both task types. In task A confusion was induced by an incongruent sentence (Durso et al., 2012) and in task B by a task differing from the other tasks of the same type. Based on the previously described theoretical basis, these discrepancies reflect an obstacle to task accomplishment and should result in an emotional reaction, in the form of confusion. Participants in the CG did not receive incongruent information or divergent tasks. During both tasks participants' facial expressions were recorded via a webcam for post-hoc analysis with the affective computing software to detect AU4. To ensure even lighting of the faces while avoiding blinding respondents, the laboratory booths were equipped with softlights. The sequence of the tasks was randomized. The study design is shown in Figure 2. Participants were paid a show-up fee of € 4, plus a payoff based on their performance in the tasks to ensure motivation (Smith, 1976). Their payoff increased with each correct answer.

Participants answered an initial questionnaire inquiring demographic data like age, gender, occupation, and

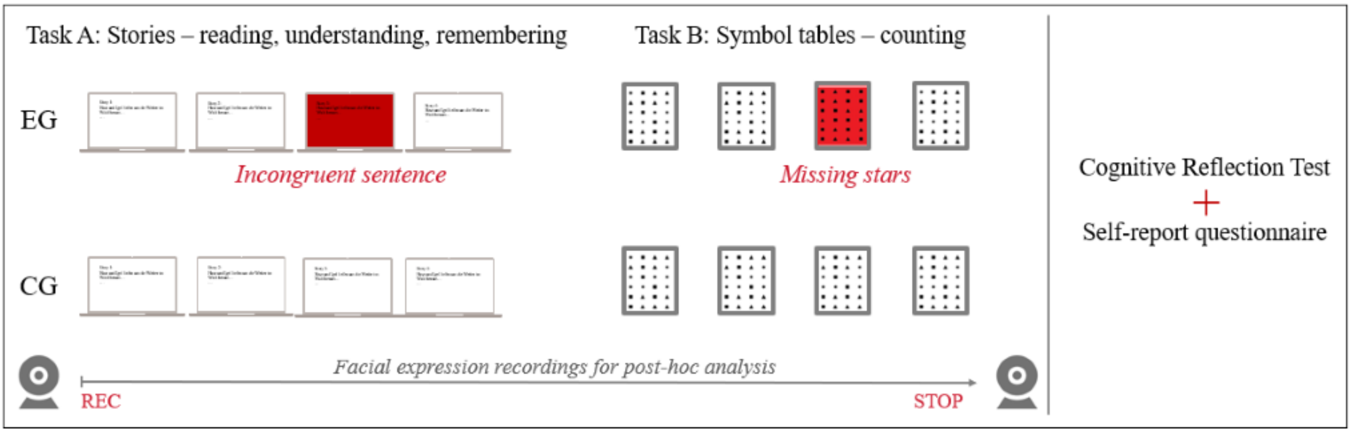

Task A: Stories – reading, understanding, remembering
Task B: Symbol tables – counting
EG
Incongruent sentence
Missing stars
CG
Cognitive Reflection Test
+
Self-report questionnaire
Facial expression recordings for post-hoc analysis
REC
STOP

Figure 2: Study design.

native language before continuing with task A or task B. Task A contained four short stories, each consisting of five sentences, adapted from a website that provides exercises for German lessons (Schaefer, nd). Sentences were presented individually on the screen one after another for eleven seconds each. After each story a multiple-choice test with three questions tested memory performance. For each correct answer participants earned € 0,30. In the EG, the fourth sentence of the third story did not fit in with the rest of the story and thus represented the independent variable. To ensure the confusion is solved (VanLehn et al., 2003), the fourth sentence of the fourth story was used. Thus, confusion was resolved when reading the subsequent story that contained the now familiar, previously incongruent sentence. In the CG, the third story did not contain an incongruent sentence.

In task B participants were exposed to four consecutive tables containing 150 symbols (triangles, squares, stars). Participants were asked to count the number of stars (target symbols) among the distractor symbols in the table within a maximum of 50 seconds (see Abeler et al., 2011). A minimum time in which participants could not click "Next" was set to 10 seconds to ensure long enough recordings to capture potential confusion expressions of all respondents. For each correct answer they earned € 0,80. In the EG, the confusion was induced by the third table that did not contain stars. Confusion was resolved when participants noticed the correct answer was "0". The third table in the CG also contained stars. The exact wording of the stories and pictures of the symbol tables can be found in the Appendix.

Following the tasks, participants were asked to solve the *Cognitive Reflection Test* (CRT; Frederick, 2005). The CRT consists of three items (see Appendix) for which most peoples' intuitive and spontaneous answers are incorrect but that can be correctly solved by cognitive effort and System 2 reflective processes. For each correct answer participants earned € 0,80. At the end of the study subjects answered a self-report questionnaire on how confusing each task (A, B, and CRT) was for them.

2.3 Analysis of video recordings

To analyze the video recordings, the software generated an AU4 intensity value for each frame (recordings were made with 15 frames per second). The AU4 values vary between 0 and 1. For each participant, eight AU4 time series were generated (representing face recordings during the four stories and the four tables). To condense the time series data, individual mean values for AU4 was computed for each sentence of each story (i.e., for 20 sentences) and for each symbol table resulting in 24 mean AU4 values for each respondent.

3 Results

3.1 Control variables and self-report

The distribution of the control variables age, gender, and occupation did not differ between the groups, so that their effect on the dependent variables can be ignored.

To test if the distribution of the answers in the self-report of the EG differs significantly from the answers of the CG, a Mann-Whitney-U-Test was computed. Whereas a highly significant difference in the perception between the EG (M = 2.84, SD = 0.98) and the CG (M = 2.06, SD = 0.89) was found for task A, W(37,34) = 901.5, $p < .001$, the distribution of the answers for the symbol task did not differ significantly between the EG (M = 2.38, SD = 1.01) and the CG (M = 2.41, SD = 1.02), W(37,34) = 622.5, p = .942.

3.2 Detection of AU4

We first checked whether the confusion manipulations indeed led to a corresponding, discernible facial reaction, namely activation of AU4. This was tested by within- and between-subjects comparisons. The within-subjects comparisons examine whether participants in the EG show higher AU4 values in the confusing part of the task (sentence 4 of story 3, and symbol table 3) compared to the parts before (sentences 1-3 of story 3, and symbol tables 1-2). The between-subjects comparisons test if the EG has higher AU4 values in the confusing part of the task (sentence 4 of story 3 and symbol table 3) than the CG has in the corresponding sections.

Task A: Stories
Figure 3 depicts the time course of AU4 values averaged across respondents for both groups while reading the third story. The fourth sentence (i.e., the confusing part for the EG) is denoted by a black frame. The curve of the averaged AU4 values of the EG increases while reading the story and reaching its peak while reading the fourth sentence.

Comparing the respondents' mean AU4 values per sentence, the one-tailed t-test for dependent samples is significant, t(36) = - 1.88, p = .034, showing that the EG had higher AU4 values while reading the fourth sentence (M = 0.07, SD = 0.10) compared to reading the three sentences before (M = 0.03, SD = 0.05). Therefore, for task A the within-subjects comparison suggests that the confusion manipulation indeed induced a facial confusion response.

Comparing the respondents' mean AU4 values for the fourth sentence between groups (i.e., mean values just for the highlighted section), the AU4 values of the EG reach a higher level than those of the CG. To test the significance of the between-subject comparison, a one-tailed t-test for independent samples was computed. The EG showed significantly higher AU4 values (M = 0.07, SD = 0.10) while reading the fourth sentence than the CG (M = 0.04, SD = 0.07), t(63) = 1.65, p = .052. Thus, also the between-subjects comparison indicates that the confusion manipulation had the intended effect.[1]

Task B: Symbol tables
Figure 4 depicts the time course of the AU4 values for the first 10 seconds of the first three symbol counting tasks, averaged across participants. Figure 4 indicates that the AU4 values of the EG are higher while working on the third symbol table than while working on the two tables before.

To compare the respondents' mean AU4 values per table, a one-tailed t-test for dependent samples was conducted. The EG shows significantly higher AU4 values in the third table (M = 0.06, SD = 0.07) compared to the two tables before (M = 0.05, SD = 0.07), t(36) = -1.73, p = .046. Thus, also for task B we can conclude from the within-subjects comparison that the confusion manipulation was successful.

Comparing the respondents' mean AU4 values between groups during the counting of the third symbol table, the AU4 values of the EG reach a higher level than the values of the CG. A one-tailed t-test for independent samples showed that AU4 values while counting table 3 were significantly higher in the EG (M = 0.06, SD = 0.07) than in the CG (M = 0.03, SD = 0.06), t(70) = 1.60, p = .058. So last but not least, the between-subjects comparison for task B shows that the experimental manipulation had the intended effect and respondents indeed expressed confusion.

Thus, in sum, participants in the EG expressed more confusion during the confusing parts of the tasks than before the confusion manipulation, and did so in both the memory and the attention task. Also, participants in the EG expressed more confusion during the confusing parts of the tasks than participants in the CG did in the corresponding task parts. Again, this holds for both task types.

3.3 Contextual performance

Hypotheses 1 and 2 posit a positive impact of confusion on contextual performance and was tested by comparing performance between the groups in the tasks that immediately followed the confusion-inducing tasks.

Task A: Stories
The percentages of participants in each group with 1, 2, or 3 correct answers in the multiple-choice test for story 4 is presented in Figure 5. It shows that more participants in the EG solved all three questions than in the CG. Statistical significance of the difference was tested by comparing the number of correct solutions in a one-tailed Mann-Whitney-U test for independent samples (U = 779.5, p = .0385) which shows that the number is significantly higher on average in the EG (M = 2.65, SD = 0.72) than in the CG (M = 2.43, SD = 0.70). This means Hypothesis 1 is supported.

Task B: Symbol tables
Descriptive statistics show that the percentage of correct answers for table 4 is higher in the CG (35 %) than in the EG (30 %). However, a Chi-squared test shows that the difference is not statistically significant, χ^2 (1, N = 72) = 1.10, p = .293. Thus, Hypothesis 2 cannot be confirmed for the symbol tables; even a slight trend in the opposite direction is revealed.

3.4 General cognitive performance

Hypothesis 3 states that confusion also has a positive effect on general cognitive performance which was measured with the CRT. Figure 6 shows the percentages of participants in each group with 0, 1, 2, or 3 correct answers in the CRT. It shows that in the CG, participants who had 0 correct answers represent the largest group (40%), while in the EG, it is participants who got one answer correct (32%). In fact, the percentages of participants with 1, 2 and 3 correct answers are all higher in the EG than in the CG. On average, 1.38 (SD = 1.06) correct answers were given in the EG, and 1.03 (SD = 1.04) correct answers were given in the CG. A one-tailed Mann-Whitney-U test for independent samples showed

[1] To acknowledge the exploratory nature of the study, we apply a significance level of $\alpha = .1$ in all statistical tests. We are aware that we are thus more lenient in rejecting the null hypothesis than is usual in experimental psychology, but point out that the applied level is often used in related disciplines (see, e.g., Cameron and Trivedi, 2005, p. 248). All in all, we think that the pattern of results is promising, but replication studies are needed to corroborate the findings.

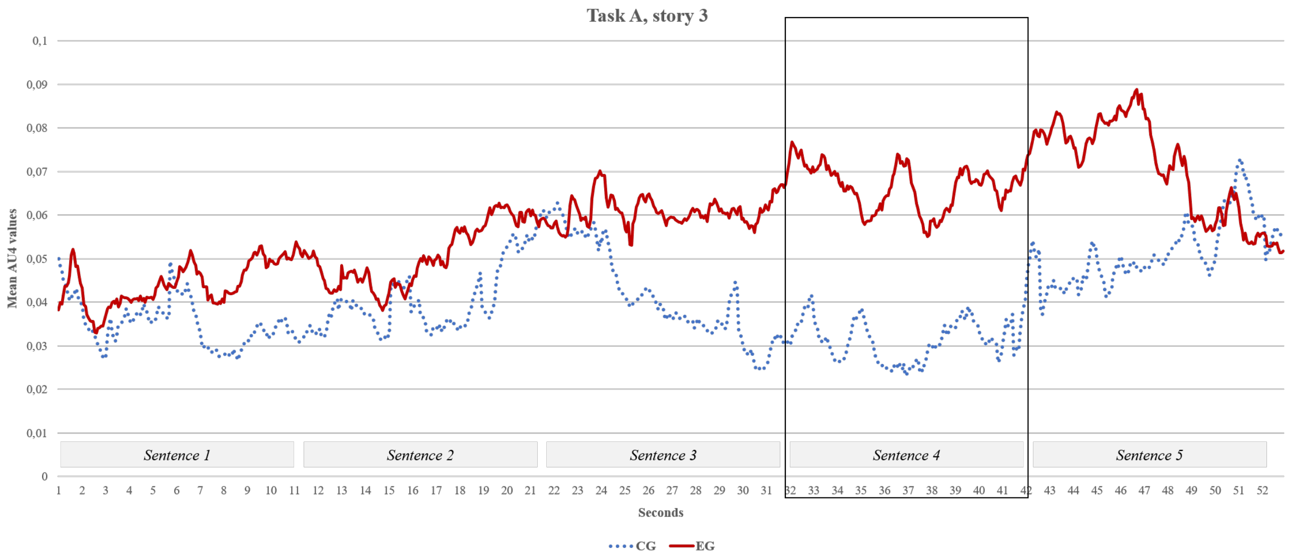


Figure 3: Line diagram of the time course of the AU4 values while reading the third story of the memory task (i.e., task A), averaged across participants for each group.

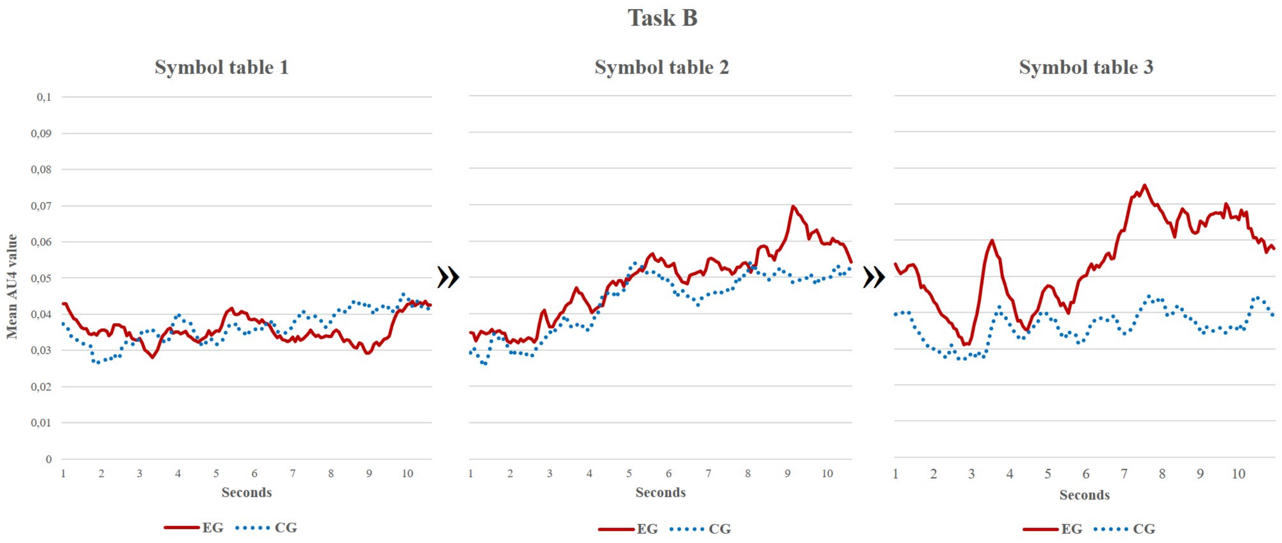


Figure 4: Line diagram of the time course of the AU4 values for the first 10 seconds of the first three tables of the attention task (i.e., symbol counting, task B), averaged across participants for each group.

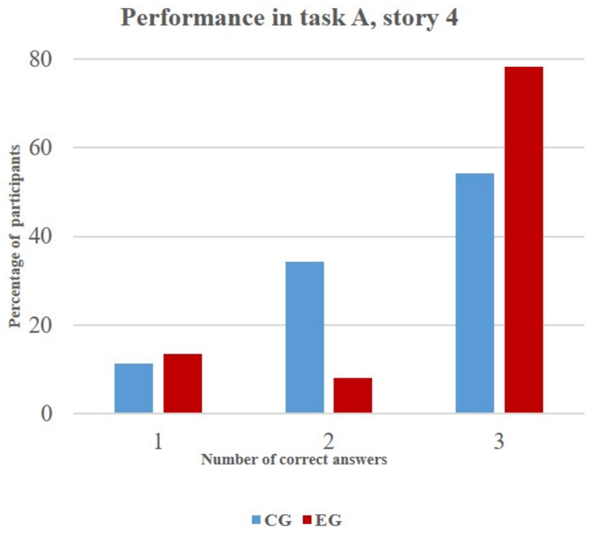


Figure 5: Percentage of participants per number of correct answers in the multiple-choice test about story 3 of the memory task.

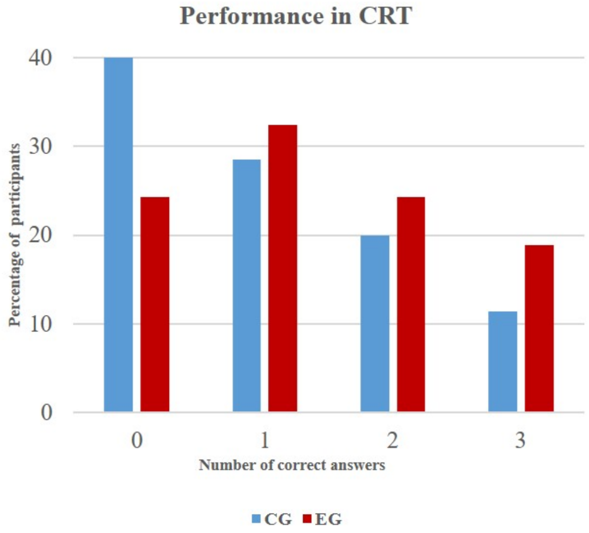


Figure 6: Percentage of participants per number of correct answers in the CRT.

that the difference is significant (U = 769.5, p = .077). Again, the hypothesis is supported.

In sum, the results suggest that confusion indeed increased contextual performance in the memory task and general performance in the CRT, but not contextual performance in the attention task.

4 Discussion

The aim of this study was to examine the effect of confusion on information processing and thinking processes. Confusion detection was based on automatic analysis of facial expressions. The software proved sensitive enough to detect confusion-induced increases in AU4 in the EG.

Only one of the experimental tasks (the memory task) was retrospectively perceived as more confusing in the EG than in the CG, which can be explained by the relatively short state of confusion in the attention task before its resolution. This also highlights the benefits of capturing emotional responses in real-time, that is, exactly at the point in time when they happen. Emotions can be brief, fleeting and not sufficiently pronounced to be consciously perceived and remembered, which may lead to underreporting in post-hoc questionnaires (Rosenberg, 1998).

For both experimental tasks, it was demonstrated that experimentally induced confusion leads to higher AU4 values, which can be captured by affective computing software. However, the effect was relatively small, especially in the between-subjects comparisons. One reason for this could be large interindividual level differences in facial expression (Cohn et al., 2002). It is also important to note that the averaged AU4 values over the whole sentences or the whole tables were used for hypotheses testing. This represents a rather conservative indicator, as shorter mimic reactions may be levelled out. Future work should aim at identifying potentially more sensitive indicators that take into account the fleeting nature of emotion expressions in the time course, such as individual maxima, variability measures, or sharp local upward slopes (D'Mello et al., 2018).

For the attention task, no effect of confusion on performance was found. One explanation could be that increased attention due to confusion is too short and fleeting to positively affect performance in the subsequent task. For a sustained benefit of confusion, it might be necessary that participants in a state of heightened attention then expend effort to deliberate, re-think, and engage in deeper learning processes. These cognitive activities may improve memory performance when reading a text, but are not required for the task of counting target symbols in a table. Future research should further address the question for which cognitive tasks exactly confusion can be beneficial.

The postulated positive impact of confusion on contextual performance was indeed confirmed for the memory task. A positive effect on general cognitive performance was confirmed as well. It is conceivable that there is a common mechanism behind both effects, namely confusion-triggered deliberate cognitive reflection, in the form of contextual deep learning in the memory task and System 2 activation in the CRT. However, the effects both for contextual performance in the memory task (d = 0.31) and for general cognitive performance (d = 0.33) were small (Cohen, 2007).

One explanation for the small effects could be that the dependent variable performance was measured with too few items, limiting the range of possible outcomes. It can be assumed that more items would have led to a more fine-grained performance differentiation between respondents, which may yield stronger effects. Furthermore, external factors likely to affect performance in the CRT, such as IQ and previous experiences with the CRT or similar tests should be controlled in future research. Finally, the small sample of 72 participants represents another limitation of the present study.

A more fine-grained operationalization of the dependent variables, controlling for external influence factors, and a larger sample should be taken into consideration when trying to replicate the results of the present study.

5 Conclusion

The findings of the present investigation provide evidence for the impact of inducing confusion during learning activities on performance in an experimental setting. Results hint at a performance-enhancing effect of confusion that is resolved briefly after the confusing event for tasks that require memorization and text comprehension. These findings could be used in educational contexts by systematically building obstacles and inducing confusion to succeed in a challenging task. Solving the obstacles can offer opportunities to intensify learning activities (Bjork and Linn, 2006; Bjork and Bjork, 2011).

The present study also gives some indication for a positive impact of confusion on general cognitive performance, which might have implications for decision-making contexts. Frank and Magnone (2011) state that "surprises are bias killers". Likewise, a temporary state of confusion due to expectancy violations and incongruencies could help avoiding mistakes and increase decision quality by acting as a cue to reflect on rash decision-making routines and to engage in more analytic and reflective processing by activating System 2 (Alter et al., 2007).

Induced confusion led to a facial expression of AU4 and can be detected by affective computing software. The automatic detection of confusion can serve as an additional information to help the confused person recognize her own confusion and use it to her advantage. In particular, the ability of the software to detect confusion even if the person herself did not perceive it can be beneficial. Given real-time capability of the automatic analysis, detected confusion could be reported back to the person at exactly the point in time when it occurs, combined with delivering hints that help recognize and ultimately overcome the obstacle. To achieve this, the detection software could be implemented in decision support systems. Besides real-time capability, a prerequisite for such an application, however, is high diagnosticity at the individual level, not just at the group level as reported in the present study. For this purpose, identification of more sensitive indicators for confusion expression than simple averaging over time appear necessary to identify confusion exactly when a person starts experiencing it, for instance, by a steep increase of the AU4 value.

Overall, the findings underline the importance of further studies on the impact of confusion on performance. More research is necessary before confusion can be strategically used for deeper information processing or for decision support. This study offers a first indication that there can be a positive impact not only on contextual performance but also on general cognitive effort.

References

Abeler, J., Falk, A., Goette, L., and Huffman, D. (2011). Reference points and effort provision. *American Economic Review*, 101(2):470–492.

Alter, A., Oppenheimer, D., Epley, N., and Eyre, R. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General*, 136(4):569–576.

Berlyne, D. (1960). *McGraw-Hill Series in Psychology. Conflict, Arousal, and Curiosity.* McGraw-Hill Book Company, New York.

Bjork, E. and Bjork, R. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In Gernsbacher, M., Pew, R., Hough, L., Pomeranzt, J., and Foundation, F., editors, *Psychology and the Real World: Essays Illustrating Fundamental Contributions to Society*, pages 56–64. Worth Publishers.

Bjork, R. and Linn, M. (2006). The science of learning and the learning of science: Introducing desirable difficulties. *The American Psychological Society Observer*, 19(3):29,39.

Brosch, T. and Scherer, K. (2009). Komponenten-Prozess-Modell: Ein integratives Emotionsmodell. In Brandstätter, V. and Otto, J., editors, *Handbuch der Allgemeinen Psychologie: Motivation und Emotion*, pages 446–456. Hogrefe.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: methods and applications.* Cambridge university press.

Chen, D., Schonger, M., and Wickens, C. (2016). oTree-An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Chinn, C. and Brewer, W. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, 63(1):1–49.

Cohen, J. (2007). A Power Primer. *Tutorials in quantitative methods for psychology*, 112(1):155–159.

Cohn, J., Schmidt, K., Gross, R., and Ekman, P. (2002). Individual differences in facial expression: stability over time, relation to self-reported emotion, and ability to inform person identification. In *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, pages 491–496.

Collins, A., Warnock, E., Aiello, N., and Miller, M. (1975). Reasoning from incomplete knowledge. In *Representation and Understanding*, pages 383–415. Morgan Kaufmann, San Diego.

Craig, S., D'Mello, S., Witherspoon, A., and Graesser, A. (2008). Emote aloud during learning with AutoTutor: Applying the Facial Action Coding System to cognitive-affective states during learning. *Cognition and Emotion*, 22(5):777–788.

Craig, S., Graesser, A., Sullins, J., and Gholson, B. (2004). Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29(3):241–250.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals.* John Murray, London. Retrieved from http://darwin-online.org.uk/.

D'Mello, S. (2013). A selective meta-analysis on the relative incidence of discrete affective states during learning with technology. *Journal of Educational Psychology*, 105(4):1082–1099.

D'Mello, S. and Graesser, A. (2011a). Confusion. In Pekrun, R. and Linnenbrink-Garcia, L., editors, *International Handbook of Emotions in Education*, pages 289–310. Routledge.

D'Mello, S. and Graesser, A. (2011b). The half-life of cognitive-affective states during complex learning. *Cognition and emotion*, 25:1299–308.

D'Mello, S. and Graesser, A. (2012). Dynamics of affective states during complex learning. *Learning and Instruction*, 22(2):145–157.

D'Mello, S., Kappas, A., and Gratch, J. (2018). The affective computing approach to affect measurement. *Emotion Review*, 10(2):174–183.

D'Mello, S., Lehman, B., and Person, N. (2010). Monitoring affect states during effortful problem solving activities. *I.J. Artificial Intelligence in Education*, 20:361–389.

Durso, F., Geldbach, K., and Corballis, P. (2012). Detecting confusion using facial electromyography. *The Journal of the Human Factors and Ergonomics Society*, 54(1):60–69.

Durso, F. and Gronlund, S. (1999). Situation awareness. In Durso, F., Nickerson, R., Schvaneveldt, R., Dumais, S., and Chi, M., editors, *The Handbook of Applied Cognition*, pages 283–314. Wiley, New York.

Ekman, P. and Friesen, W. (1978). *Facial Action Coding System: A Technique for the Measurement of Facial Movement.* Consulting Psychologists Press, Palo Alto.

Festinger, L. (1957). *A Theory of Cognitive Dissonance.* Consulting Psychologists Press, Palo Alto.

Forbes-Riley, K. and Litman, D. (2009). Adapting a student uncertainty improves tutoring dialogues. In Dimitrova, V., Mizoguchi, R., and Boulay, B. D., editors, *Proceedings of 14th international conference on artificial intelligence in education*, volume 200, pages 33–40, Amsterdam. IOS Press.

Forbes-Riley, K. and Litman, D. (2011). Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Communication*, 53(9-10):1115–1136.

Forbes-Riley, K. and Litman, D. J. (2010). Designing and evaluating a wizarded uncertainty adaptive spoken dialogue tutoring system. *Computer Speech and Language*, 25(1):105–126.

Frank, C. and Magnone, P. (2011). *Drinking from the Fire Hose: Making Smarter Decisions Without Drowning in Information.* Portfolio, New York.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19(4):25–42.

Graesser, A., Chipman, P., King, B., McDaniel, B., and D'Mello, S. (2007). Emotions and learning during autotutor. In Lucking, R., Koedinger, K., and Greer, J., editors, *Proceedings of the 13th international conference on artificial intelligence in education*, pages 569–571, Amsterdam. IOS Press.

Graesser, A. and Olde, B. (2003). How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down. *Journal of Educational Psychology*, 95(3):524–536.

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

Laird, J., Newell, A., and Rosenbloom, P. (1987). Soar - an architecture for general intelligence. *Artificial Intelligence*, 33(1):1–65.

Lehman, B., Matthews, M., D'Mello, S., and Person, N. (2008). What are you feeling? Investigating student affective states during expert human tutoring sessions. In Woolf, B., Aïmeur, E., Nkambou, R., and Lajoie, S., editors, *Intelligent Tutoring Systems*, pages 50–59. Springer Berlin Heidelberg, Berlin.

Mandler, G. (1975). *Mind and Emotion.* Wiley, New York.

Mandler, G. (1984). *Mind and Body: Psychology of Emotion and Stress.* W.W. Norton and Company, New York.

Mandler, G. (1990). Interruption (discrepancy) theory: Review and extensions. In Fisher, S. and Cooper, C., editors, *On the move: The psychology of change and transition*, pages 871–890. Wiley, Chichester.

Piaget, J. (1952). *The Origins of Intelligence in Children.* W.W. Norton and Company, New York.

Richey, J., Andres-Bray, J., Mogessie, M., Scruggs, R., Andres, J., Star, J., Baker, R., and McLaren, B. (2019). More confusion and frustration, better learning: The impact of erroneous examples. *Computers and Education*, 139:173–190.

Rodrigo, M., Baker, R., Jadud, M., Amarra, A., Dy, T., Espejo-Lahoz, M., Lim, S., Pascua, S., Sugay, J., and Tabanao, E. (2009). Affective and behavioral predictors of novice programmer achievement. In *Proceedings of the ACM SIGCSE Annual Conference on Innovation and Technology in Computer Science Education*, volume 41, pages 156–160.

Rosenberg, E. (1998). Levels of analysis and the organization of affect. *Review of General Psychology*, 2(3):247–270.

Rozin, P. and Cohen, A. (2003). High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of Americans. *Emotion*, 3:68–75.

Schaefer, S. (n.d.). [Text comprehension exercises]. Retrieved from http://suz.digitaleschulebayern.de/index.php?id=16/.

Schneider, B., Krajcik, J., Lavonen, J., Salmela-Aro, K., Broda, M., Spicer, J., Bruner, J., Moeller, J., Linnansaari, J., Juuti, K., and Viljaranta, J. (2015). Investigating optimal learning moments in U.S. and Finnish science classes. *Journal of Research in Science Teaching*, 53(3):400–421.

Seuss, D., Hassan, T., Dieckmann, A., Unfried, M., Scherer, K., Mortillaro, M., and Garbas, J. (2020). A hybrid Gaussian-based fusion approach to Action Unit detection for inference of emotional appraisals. Manuscript submitted for publication.

Smith, V. (1976). Experimental economics: Induced value theory. *The American Economic Review*, 66(2):274–279.

VanLehn, K., Siler, S., Murray, C., Yamauchi, T., and Baggett, W. (2003). Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21(3):2009–249.

A Appendix

A.1 Task A, Story 3

1. Als Julia mit Kind und Kegel das Haus verließ, befanden sich entlang der Straße einige Dorfbewohner, um sich von ihr zu verabschieden und ihr alles Gute auf der langen Reise nach Deutschland zu wünschen.

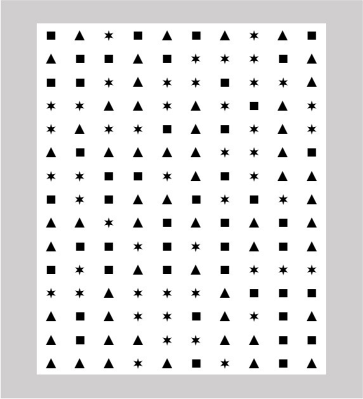
2. Dort sollte sie bei ihren Großeltern mütterlicherseits leben, bis sie in der Nähe eine Bleibe finden würde.

3. Außer ihrer Tante, deren Ehemann und dessen Eltern wohnten in derselben Straße die Großeltern väterlicherseits, die ein paar Haustiere hatten: Hunde, Katzen, Hasen und weiße Mäuse.

4. ***CG:*** *Binnen weniger Tage fühlten sich die Kinder in ihrem neuen Zuhause sehr wohl, denn sie hatten ihr eigenes Zimmer und trafen gerne ihre Cousins und Cousinen.*
   EG: *Als es aber sah, dass sich die Hühner untereinander ebenso bissen, war es beruhigt und ertrug die Feindseligkeiten mit Gleichmut.*

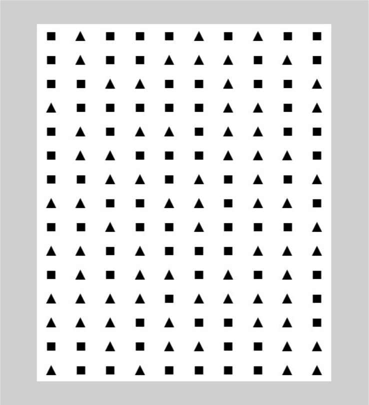
5. Außerdem gab es vor dem Haus einen Spielplatz, auf dem sich die Kinder gerne aufhielten.

A.2 Task B, Table 3



■	▲	✶	■	▲	■	▲	✶	▲	■
▲	■	■	▲	■	✶	✶	✶	■	▲
■	■	✶	▲	✶	✶	■	✶	✶	▲
✶	✶	▲	▲	✶	▲	✶	■	▲	✶
✶	▲	✶	✶	■	▲	■	✶	▲	✶
▲	■	▲	▲	▲	▲	✶	✶	▲	■
✶	✶	■	■	✶	▲	■	✶	▲	▲
■	✶	■	▲	▲	■	✶	■	✶	▲
▲	▲	✶	▲	■	▲	■	▲	■	▲
▲	■	■	✶	■	✶	■	▲	■	▲
■	✶	■	▲	■	▲	■	✶	✶	✶
✶	✶	▲	✶	✶	✶	▲	■	■	■
▲	■	▲	✶	✶	■	▲	✶	■	▲
▲	■	▲	▲	✶	✶	▲	▲	■	■
▲	▲	▲	✶	▲	■	✶	▲	■	▲

Figure 7: Table 3 presented to the CG.



■	▲	■	■	■	▲	■	▲	■	■
■	▲	■	■	▲	▲	▲	■	▲	■
■	■	▲	▲	■	■	▲	■	■	▲
▲	■	■	■	■	■	▲	▲	■	▲
■	▲	■	▲	▲	■	▲	▲	■	■
■	▲	▲	■	■	■	▲	▲	▲	■
■	■	▲	▲	■	▲	■	▲	■	▲
▲	▲	■	■	▲	▲	■	▲	▲	■
■	■	▲	■	■	▲	■	■	■	▲
▲	▲	■	▲	■	■	■	▲	▲	▲
■	▲	■	▲	▲	■	▲	■	▲	■
▲	▲	▲	▲	■	▲	▲	▲	▲	■
▲	▲	▲	■	▲	■	■	▲	▲	■
■	■	▲	■	▲	▲	■	■	▲	▲
▲	■	■	▲	■	■	■	■	▲	▲

Figure 8: Table 3 presented to the EG.

A.3 Cognitive Reflection Test

The CRT consists of the following three questions:

1. A bat and a ball cost \$ 1.10 in total. The bat costs \$ 1.00 more than the ball. How much does the ball cost?

2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

The correct answers are: 5 cents, 5 minutes, and 47 days.