**Grantee:**

Djellel Eddine Difallah
University Of Fribourg, Switzerland

**Host:**

Dr. Gianluca Demartini
University of Sheffield, UK

*Proposal Title:*

Measuring Human Behaviour in Search Evaluation Micro Tasks.

*Application Reference N°:*

7091

1. **Purpose of the visit**

Crowdsourcing is becoming a viable alternative to scale the generation of relevance judgements to evaluate information access systems. However, more than simple relevance judgements, user-oriented interactive evaluation or IR systems require user-focused studies. During this short visit, we propose to investigate crowdsourcing as a means to scale-out user-entered studies.

Some studies might rely on crowdsourcing more than others, or simply have much more more input. It is not unusual to see hundreds of thousands of similar tasks published from a single source. For a crowd worker this means long hours of repetitive tasks, e.g., judging relevance.

While the effects of repetitive manual tasks are well studied in psychology, those results cannot be directly applied or assumed for online micro tasks.

In the proposed visit, we will be collaborating with researchers at the University of Sheffield both from the Information School (i.e., the Information Retrieval group) and the Department of Psychology. The goal of the visit is to layout a measurement framework of *Flow Experience*, that is, the mental state a worker can attain if fully immersed in the given tasks. The main question to investigate is whether it is possible to model the user behaviour from historical data and clearly distinguish the three states of Flow theory: boredom, flow and anxiety.

We plan to meet experts in psychology to design experiments for better understanding worker behaviour in crowdsourcing platforms. By understanding in how workers feel in different situation, it will be possible to design more reliable evaluation tasks.

## 2.    Description of the Work Carried out During the Visit

### 2.1.    Background and Related Work

In the field of psychology, the concept of Flow[1] is well established. As Figure 1 illustrates, being in the flow state is a balance between the skills that the person has in conducting an activity with a certain difficulty level. As such, if the person is overskilled, i.e., has high skills as compared to the given task, he might quickly get bored. Likewise, if the person does not have the skills to conduct a complex task, he might quickly get anxious.  Flow is usually measured using the Flow Short Scale[2]; a questionnaire that is proposed to the participants at the end of an experiment. However, we did not find any relevant work studying the flow theory as applied to online work (crowdsourcing), nor evidences that this concept applies to work done on current crowdsourcing platforms such as Amazon Mechanical Turk[3].
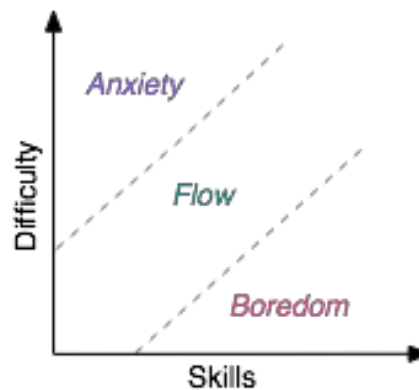


Figure 1 - The Concept of Flow Theory.

### 2.2.    Developed Idea

Given the repetitive and potentially dull nature of micro-tasks, a batch of HITs can be dynamically altered so that it could continuously challenge or relax the worker to keep him/her in a Flow state. The system should automatically sense and act respectively in order to help the worker reach and maintain that state.

Our research direction is to create a strategy based on the expected response time of each task type. For example, if the worker is exhibiting a response time that is consistently lower than the mean, then the worker might be too skilled for the task at hand and can eventually get bored. On the contrary, if the response time is  higher than the mean, then the worker is most probably struggling with the task. The system will then dynamically respond to these signals by proposing easier or respectively more challenging tasks.

---

[1] Csikszentmihalyi, Mihaly, and Mihaly Csikzentmihaly. *Flow: The psychology of optimal experience*. Vol. 41. New York: HarperPerennial, 1991.

[2] Rheinberg, Falko, Regina Vollmeyer, and Stefan Engeser. *Die Erfassung des Flow-Erlebens*. na, 2003.

[3] http://www.mturk.com

## 2.3. Experimental Design

We start with the hypothesis that maintaining *Flow* is desirable for a micro-task worker; both by improving his/her experience, and also maintaining a high answer quality and low response time. If we are able to keep the workers in their Flow zone, we can contribute to more engaged and productive workers.

Our research methodology will be quantitative and qualitative, hence, through a well thought set of experiments we will explore the Flow concept in crowdsourcing. We start by defining the following aspects in our intended study.

### 2.3.1. Task Type and Difficulty

The complexity of a task at hand in a crowdsourcing setup can be varied using different methods. For instance, we can alternate the task types based on their intrinsic complexity e.g., from image tagging, to audio transcription to Internet search etc. We can also vary the task difficulty of the same type e.g., audio transcription of poor recordings. Alternatively, we can artificially challenge the worker by setting a timer on a given task or by changing the reward value based on his/her previous performance.

In order to focus our study, we decided to limit our experiments on a single task type, that is finding image duplicates[4] (see Figure 2 for a screenshot). The user is presented with an image on the left with which he is tasked to find a duplicate among a set of images on the right. For scalability purposes, instead of opting for a task with real images (or text), we will generate images to be matched automatically using an image generation library.



Figure 2 - The task matching User Interface. Find a duplicate of the left image with the images shown on the right.

The difficulty is then controlled by changing the number of images on the right grid (see Figure 3). We have decided to experiment with a few difficulty strategies a) A fixed difficulty level to capture static data, b) ask the user to change the difficulty manually to capture boredom or

---

[4] An example of the developed software can readily be found on: http://xi-lab.github.io/inFlux/etol/

anxiety, c) Increasing the difficulty gradually (using various interval levels), d) interleave difficult tasks with easy tasks to simulate work breaks. The final goal would be to create an engine that changes the difficulty dynamically based on a Flow model that we would develop based on the results of these strategies.



Figure 3 - different the difficulty levels for our image duplicate finding task.

In order to be fair, each task difficulty will have its own price reward; the more pictures to match the higher the reward. It is however not clear that the reward needs to be linearly proportional to the number of tasks. For that, we performed an ETA[5] analysis of the price at discrete difficulty level and interpolate (a recent technique to evaluate the fair reward of a task). In an experiment we ran, we asked our study participants to complete 200 tasks containing respectively 6, 24 and 60 images to match. The results of the ETA analysis are shown in Figure 4, where the results indicate that to be within 10% error rate the workers need: 3.88, 4.9, 5.86 seconds respectively. Hence, for an hourly rate of $12/hour the task rate should be $2.7, $3, $4 respectively.
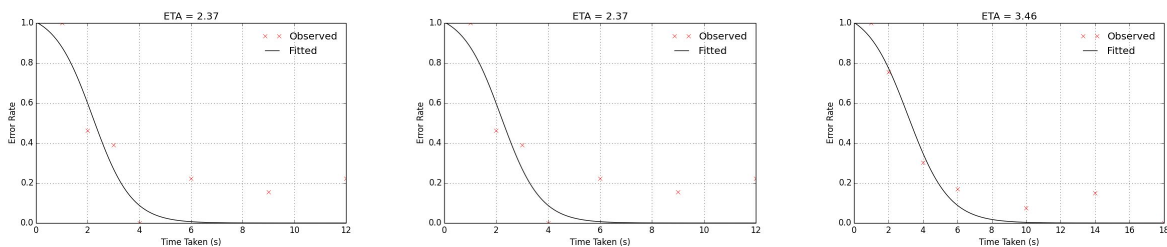


Figure 4 - ETA for the image duplicate finding task. Results from left to right represent ETA for 6, 24 and 60 images respectively.

### 2.4.    Evaluation

In the previous section, we have defined the task type and the difficulty dimension for the flow theory. What we need now is the observed dimension, that is the *"skill"*. In our case, we initially

---

[5] Cheng, Justin, Jaime Teevan, and Michael S. Bernstein. "Measuring Crowdsourcing Effort with Error-Time Curves." *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 2015.

defined the skill as a ratio of the quality of the answers (correctness) and the response time (how fast was the answer submitted). The idea is that the workers will develop specific skills at identifying duplicates with larger number of candidates. In addition to correctness and response time, we ask the study participants to fill out the standard flow questionnaire at the end of an experiments (see Figure 4).



Figure 5 - Flow Measurement.

### 3. Future Collaboration with Host Institution

Crowd-powered systems present unique challenges and opportunities especially in the context of scalability. This work goes along my work on worker retention to lower execution time. While in a previous work I explored the effect of pricing on retention, this work is a new take on retention using flow theory. During my visit I had the chance to collaborate with Dr. Demartini who introduced me to Dr. Tom Stafford from the department of psychology of the University of Sheffield who gave us great insight and feedback on how to carefully design psychological experiments and how to interpret flow measurements.

### 4. Other Comments

With this visit we started a new scientific endeavor that will be carried out during the next 6 months. The final outcome and the scientific findings that will result from this work will be submitted to relevant conferences and/or journals for peer reviewing. Upon publication (including technical reports) I will gladly acknowledge ESF for their fundings and support.