Microwork platforms as enablers to new ecosystems and business models: the challenge of managing difficult tasks

Jean-Michel Dalle*

Université Pierre et Marie Curie,
Paris, France
and
CRG-i3,
Ecole Polytechnique, France
Email: jean-michel.dalle@upmc.fr
*Corresponding author

Matthijs den Besten

Montpellier Research in Management,
Montpellier Business School,
Montpellier, France
Email: m.den-besten@montpellier-bs.com

Catalina Martínez

Institute of Public Goods and Policies (CSIC-IPP),
Madrid, Spain
Email: catalina.martinez@csic.es

Stéphane Maraut

Madrid, Spain
Email: stephane.maraut@gmail.com

Abstract: We explore how microwork platforms manage difficult tasks in paid crowdsourcing environments. We argue that as human computation becomes more prevalent, notably in the context of big data ecosystems, microwork platforms might have to evolve and to take a more managerial stance in order to provide the right incentives to online workers to handle difficult tasks. We illustrate this first through a name disambiguation experiment on Amazon Mechanical Turk (AMT), a well-known microwork platform, and second through direct analysis of the dynamics of task execution in a dataset of real microwork projects on AMT. We discuss the emergence of more specialised microwork platforms as an attempt to facilitate a better management of difficult tasks in the context of paid crowdsourcing.

Keywords: crowdsourcing; online labour markets; human computation; microwork platforms; task difficulty.

Reference to this paper should be made as follows: Dalle, J-M., den Besten, M., Martínez, C. and Maraut, S. (2017) 'Microwork platforms as enablers to new ecosystems and business models: the challenge of managing difficult tasks', *Int. J. Technology Management*, Vol. 75, Nos. 1/2/3/4, pp.55–72.

Biographical notes: Jean-Michel Dalle is a Professor at University Pierre et Marie Curie (Paris 6) and an Associate Researcher with i3 (CNRS and Ecole Polytechnique). His research interests deal with the economics and management of innovation and notably with the new organisational forms and processes allowed by the internet and associated technologies.

Matthijs den Besten is an Assistant Professor of Technology and Innovation Management at Montpellier Business School. His research interests range from freelancing and entrepreneurship to new forms of collaboration permitted by the internet.

Catalina Martínez is a Research Fellow at the Institute of Public Goods and Policies in the Spanish National Research Council (CSIC). Her research interests lie in the field of science and technology policy and economics, with a focus on the economics of patents and innovation, technology markets and science-industry linkages.

Stéphane Maraut is an Independent Researcher and IT consultant based in Madrid. His research interests are related to data harmonisation, record linkage and disambiguation and he works on a broad range of projects as data architect and data scientist.

This paper is a revised and expanded version of a paper entitled 'Crowdsourcing the Names-Game: a prototype for name disambiguation of author-inventors' presented at 14th International Society of Scientometrics and Informetrics Conference (ISSI), Vienna, Austria, 15–19 July 2013 and of a paper entitled 'A direct empirical investigation of the determinants of online labour supply in Amazon Mechanical Turk' presented at the 3rd IPP Conference, Oxford, UK, 25–26 September 2014.

1 Introduction

In data-rich ecosystems (Rong et al., 2013), humans are now commonly enlisted to help with data processing. Indeed, some problems are more difficult to solve with computers alone in some settings, due to the nature of data, its variety, and/or its heterogeneity: humans are therefore assigned micro-tasks in areas where they perform better (Lewis et al., 2013). For instance, the ability of human eyes to distinguish letters and numbers in images can make human processing more efficient than algorithmic computation as in the well-known CAPTCHA and reCAPTCHA techniques (von Ahn et al., 2003; von Ahn et al. 2008, respectively).

These techniques, in which 'a machine performs its function by outsourcing certain steps to humans' is known as human-based computation, human computation or human computing.[1] It is now generally associated with the notion of *microwork* (e.g., Irani, 2015), that is, work where tasks are extremely granular or, as Lehdonvirta (2016) puts it, "work consisting of the remote completion of small information

processing tasks, such as transcribing a snippet of hand-written text, classifying an image, or categorizing the sentiment expressed in a comment".

Microwork, and specifically paid microwork, is commonly implemented with online platforms acting as intermediaries. Microwork platforms enable the hiring of human labour online by leveraging internet technologies. Their existence has considerably reduced the costs of managing large numbers of tasks and agents. Consequently distributed work can be carried out at unprecedented scale. Like other platforms (e.g., Curchod and Neysen, 2009; Boudreau, 2010; Ceccagnoli et al., 2012; Gawer, 2014; Thomas et al., 2014), paid microwork platforms adopt a many-to-many architecture in order to serve as intermediaries between sellers and buyers: in this specific case, between providers of tasks who demand online labour and online workers who supply such labour.[2]

The common perception of these platforms as allowing 'crowd work' (e.g., Vakharia and Lease, 2013; Bergvall-Kåreborn and Howcroft, 2014) is therefore somewhat misleading. The 'crowd' is, in reality, composed of many workers whose income totally or partially depends on their activity on the platforms, and the economic function of these intermediaries is to organise online labour around them. Furthermore, contrary to Boudreau and Lakhani (2013)'s initial intuition that "the management challenges in exploiting *spot* labor markets are minor compared with those in other forms of crowdsourcing", major concerns have been voiced with respect to the quality of work harnessed from microwork platforms and multiple approaches towards their resolution have been proposed.

We argue, following Acemoglu et al. (2014), that, among the many issues associated with microwork platforms, the management of difficult tasks deserves particular attention not only because it requires skilled if not expert work, but also since the ex-ante assessment of difficulty among a set of tasks is in itself non-straightforward (Section 2). We illustrate the latter issue in the context of the oldest and best-known microwork platform, Amazon Mechanical Turk (AMT), and for a set of tasks related to name disambiguation, where the notion of difficulty could be partly observed directly (Section 3). We then show that the management of difficult tasks on microwork platforms is also affected by self-selection, i.e., that online workers may self-select against difficult tasks in favour of more repetitive microwork (Section 4). We interpret as a potential remedy in this context the evolving specialisation of microwork platforms, thanks to various selection, training, reputation and incentive mechanisms, with the limit that the ability of microwork platforms to attract skilled (if not expert) workers is structurally affected by the nature of the contracts offered, an evolution partly anticipated by some platforms. Future work should notably pay attention at the effect of this evolution on the price of microwork and therefore on the demand for microwork (Section 5).

2 Microwork quality and task difficulty

Our focus here is on online labour platforms that manage virtual tasks[3] and specifically microwork. Among the variety of microwork platforms, the most prominent and earliest example is AMT. Most recent research has dealt with AMT,[4] as many data scientists have embraced AMT for a wide variety of activities, ranging from data collection (e.g., Snow et al., 2008) and image analysis (e.g., Maisonneuve and Chopard, 2012) to

interview transcription (e.g., Marge et al., 2010) or copy-editing (e.g., Bernstein et al., 2010). AMT has been heralded as a quick and easily accessible means for doing behavioural experiments (Mason and Suri, 2012). Rand (2012) reviews a number of replication studies and draws the conclusion that AMT is reliable as a platform to run experiments on. The quality of work handled on AMT is a major concern to many, however, and various strategies have been devised to improve it. As Acemoglu et al. (2014) put it, we "now have immediate access to an unlimited supply of labor and a wide pool of talent and skills, but extracting the good from the bad and managing this pool of workers is fraught with difficulties".

A well-known strategy allows requesters – those who post microwork on AMT – to require that workers pass a qualification. Alonso and Mizzaro (2012) find that workers who have passed a test are more likely to complete the tasks than other workers, whereas only a limited number of workers typically complete all tasks in a set (Bernstein et al., 2010). Furthermore, Wang et al. (2012b) find that the workers who have passed the qualification tests deliver work of slightly higher quality. In addition to prescreening, Chandler et al. (2014) suggest to select workers gradually on the basis of past engagement.

There are several types of qualification tests. The most prevalent one is to test for past performance in terms of proportion of work previously approved. Ipeirotis (2010a) observes that this test is very easy to trick. Other tests concern (self-reported) skills and the location of workers derived from their IP address. Demartini et al. (2012) find that while in general Indian workers performed worse than their US counterparts, for items related to local Indian news they performed better. In 2012, AMT added a 'master' qualification to the menu of tests that can be set. AMT attributes this qualification to workers it considers trustworthy. Restricting the tasks to 'master' workers will likely lead to higher quality results (Ipeirotis, 2012), but there is some opacity about attribution criteria. Note also that Amazon charges a higher fee for work carried out this way and that, as for instance reported by Chandler et al. (2014), 'master' workers tend to be slower to react.

Conversely, and somewhat counter-intuitively, there is evidence that for at least some kind of tasks, there is little effect of the price of individual tasks on the quality of work obtained (Mason and Watts, 2010; Mason and Suri, 2012), though, according to Horton and Chilton (2010), a higher level of effort can be expected in return for a higher pay. Hirth et al. (2013) blame the relative anonymity of workers in combination with an appeal limited to the profit motive. In order to improve quality, Shaw et al. (2011) find that it helps to indicate that payment will be linked to the extent in which individual responses conform to those given by peers.

Redundancy in responses is also a widespread strategy to counter cheating, which, according to Eickhoff and de Vries (2013), has become more prevalent. Kittur et al. (2008) found a significant increase in the quality of the data obtained after the inclusion of additional questions with verifiable answers. If answers can be verified automatically this can be used for immediate feedback, otherwise to identify misbehaviour ex post (Shaw et al., 2011).

Among other solutions to improve quality, Ipeirotis (2010b) recommends announcing the rules of the game clearly in the task description and threatening with sanctions if deficiencies are observed. Franklin et al. (2011) warn however that refusing to pay ex post may provoke a backlash from workers, who rate requesters on dedicated tools and forums such as Turk Opticon or Turker Nation. Finally, needless to say, the quality

of task formulation strongly influences the quality of results obtained in AMT (Kittur et al., 2008).

In this context, the issue of managing difficult tasks in a microwork environment has received only limited attention. When it has received attention, then mostly as part of a wider investigation into quality assurance, and notably with respect to selecting and training workers, such as in Gottlieb et al. (2014). A measure of difficulty is used by Toomin et al. (2011) but only as a control variable in the context of an experiment addressing preference measurement. Mason and Watts (2010) make use of difficulty levels for the tasks they study and observe that the number of completed tasks decreases with increasing difficulty but, since difficulty does not change the effects with regard to payment incentives, they focus on results 'averaged over difficulty levels'. Several works are also concerned with complex tasks, i.e., larger tasks suitable to be decomposed in various simpler ones, difficulty being often confounded with complexity. Noronha et al. (2011), for instance, introduce a framework for nutritional analysis and stress that "complex tasks like this are hard problems for crowdsourcing, as workers may vary drastically in experience and reliability", therefore, they "propose a workflow in which the overall problem is decomposed into small, manageable, and verifiable steps". Other examples that address complex tasks by allowing for an improved management of their 'micro-workflow' include Kittur et al. (2012), Kulkarni et al. (2012) and Ahmad et al. (2011) and open-source tools have also been developed to facilitate decomposition (Kittur et al., 2011).

The real issue with task difficulty, however, is that difficulty is generally unknown *ex ante*. Besides, more or less difficult tasks of a similar nature are often mixed, without their individual difficulty being apparent. Difficulty would only become apparent if the same tasks were performed by unskilled and by skilled solvers. Rendering tasks more granular and simpler in appearance, typically through 'closed' form presentation such as yes/no questions, might actually mislead unskilled solvers and induce them into errors that go without noticing. The introduction of redundancy is nota solution either, contrary to a common but erroneous interpretation of 'wisdom of the crowd' effects à la Hong and Page (2004), according to which a diversity of problem-solvers can give them an advantage over experts: for the advantage thus given is in searching a solution space, which is not a relevant framework for apparently simple and granular microwork tasks. In most microwork situations, the mere aggregation of individual choices cannot lead to a better solution than one reached by skilled workers and/or trained experts in the subject matter.

Actually, the first aspect of this issue – unknown ex ante difficulties – is explicitly recognised and the starting point of Acemoglu et al. (2014), although in a different context from microwork. As they focus on innovation through crowdsourcing, Acemoglu et al. (2014) do not have to deal with the second aspect of the difficulty dilemma – unnoticed errors – since, when it comes to innovation, insufficiently skilled workers are unable to finish difficult tasks and hence disclose their skill levels. This disclosure in turn allows them to suggest potential remedies such as dynamic pricing mechanisms.

3 Difficult tasks in the context of name disambiguation

To illustrate the limitations of microwork platforms such as AMT when dealing with difficult tasks, we focus first on matching names of patent inventors to names of scientific authors to identify 'author-inventors'. Matching names in large bibliographic databases is precisely non-trivial since names are often misspelt and the same person can be referred to in a variety of ways, notwithstanding the existence of homonymy. The difficulty lies in deciding whether different works with similar author names belong to the same person, or not: although automated methods are now often selected for matching and disambiguation due to limited resources (e.g., Smalheiser and Torvik, 2009; Cuxac et al., 2012; Gurney et al., 2012; Wang et al., 2012a), manual matching is considered to yield higher levels of accuracy (Veve, 2009).

In our experiment, AMT workers were asked to replicate manual checks done in the process of building a database of Spanish author-inventors described in Maraut and Martínez (2014)[5], who combined automated matching techniques with expert validation of dubious matches to identify more than 4,000 Spanish author-inventors.[6] Nine different batches of either ten or fifty distinct tasks were launched in the same week (April 2014). The tasks were randomly drawn from 99 randomly selected clusters of patents and publications likely to correspond to the same person, according to name similarity, affiliation, discipline, etc., covering in total 2,106 distinct patent-paper pair comparisons. The batches were posted on the AMT platform at randomly selected times of the day, with at least twenty hours difference. Each batch remained posted 24 hours and was visible only to workers showing good track-records, i.e., whose required qualification was to have a greater than 98% approval rate for more than one hundred previously approved tasks. Each task consisted in reviewing a cluster of patents and papers likely to correspond to the same person, according to name similarity, affiliation, discipline, and other disambiguation variables.[7] We fixed a reward of USD 0.20 per task and allowed each task to be completed by a maximum of five workers.[8]

Forty five different workers participated in the experiment; fourteen worked in more than one batch (31%); 727 tasks were completed in total. As generally reported in similar contexts, the distribution of effort was highly uneven.[9] Tasks differed in terms of number of documents (articles and patent applications) presented to workers as being potentially authored by the same person, ranging from small tasks with only two documents to the largest tasks with 61 documents, with an average number of documents per batch per task between 8 and 18 and high standard deviation. Since more than one worker could complete the same task in a given batch (up to the maximum of five workers allowed), there was more than one response for most tasks, even for the tasks offered in only one batch, and no task was systematically ignored by workers. Responsiveness of workers improved over time in terms of the proportion of a batch completed after 24 hours[10] and also, in contrast to the findings reported in other studies, weekends seemed to be worse than weekdays in terms of workers' activity.[11]

To assess the average quality of the work carried out, we calculated the rate of agreement at the pair level with the expert reviewers used by Maraut and Martínez (2014). Unsurprisingly, mean agreement rates for the 11,918 judgements on the 2,106 pairs presented to AMT workers based on ANOVA tests (not reported here), turned out to be significantly different across batches, tasks and workers. Many factors, such as size of the batch, difficulty of the tasks, number, skills and motivation of responding workers, could cause these differences.

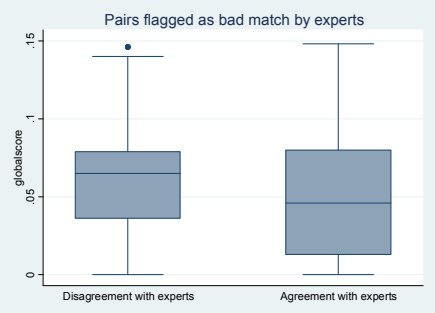
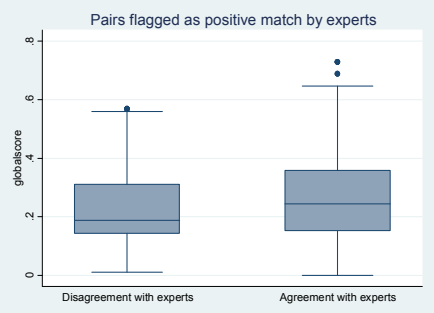
In order to test the influence of task difficulty on agreement rates, we defined a pair as 'difficult to disambiguate' making use of the variable *global score* from Maraut and Martínez (2014). In their setting, the global score is a standardised weighted combination of the values of name matching, direct disambiguation and indirect disambiguation variables at the pair level. Basically, the higher the global score, the more a pair is likely to be a positive match. Based on this, we consider that a pair is difficult to disambiguate *either* when:

1 the global score is high, but expert reviewers conclude that it is a negative match (e.g., two people with similar rare names working in the same institution)

2 the global score of the pair is low, but expert reviewers conclude that it is a positive match (e.g., a person with a common name publishing with different affiliations).[12]

Figure 1 suggests that agreement between AMT workers and experts tends to be significantly lower when pairs are more difficult to disambiguate. In other words, online workers on AMT agreed more often with experts on obvious pairs, than on more difficult ones, for both categories defined above, i.e., difficult bad matches (with high global score) and difficult positive matches (with low global score).

Figure 1 Rates of agreement with experts for bad and positive matches (see online version for colours)

Bad matches (pairs considered as such by expert reviewers)	**Positive matches** (pairs considered as such by expert reviewers)
Workers tend to disagree with experts on bad matches the higher is the global score	Workers tend to disagree with experts on positive matches the lower is the global score
They tend to disagree on pairs that looked good but were not good matches	*They tend to disagree on pairs that looked bad but where good matches*
	

An exploratory logit regression at the pair level confirms this finding. We estimate the effect of two explanatory variables on the agreement between online workers and experts: global score (continuous variable increasing with the likelihood to find a positive match based on automatic techniques) and expert validation (categorical variable equal to one when the pair had been identified as a positive match by experts and zero otherwise). We also add an interaction term, by multiplying both variables, to assess whether the effect of global score is different depending on the result of the expert validation. Furthermore, we add workers' fixed effects to control for each worker's characteristics. As shown in Table 1, when no interaction term is added, global score and expert validation have a

positive effect on agreement, but when the interaction term is included, agreement between AMT workers and experts is more likely when expert validation does not contradict the value of the global score, i.e., for the less difficult cases. Such cases would happen in two situations: when pairs have a high global score and are validated by experts (the interaction term is positive, i.e., the effect of global score when expert validation equals one); and when pairs have a low global score and have not been validated by experts (the main effect of global score is negative, i.e., the effect of global score when expert validation equals zero). We therefore argue that when global score and expert validation go in the same direction, it is easier for the worker to agree with the expert because the task is not as difficult as when they go in opposite directions.

Table 1 Logit regression

Dependent variable: agreement with experts at the pair level			
	(1)	*(2)*	*(3)*
Global score (continuous variable, takes higher values the more likely it is to be a positive match based on automatic techniques)	0.460*** (0.038)	0.162*** (0.0477)	-0.930*** (0.249)
Expert validation (dummy equal 1 when reviewer considered positive match, equal to 0 when negative)		0.171*** (0.019)	0.081*** (0.025)
Interaction global score and expert validation (global score × expert validation)			1.135*** (0.253)
AMT worker fixed effects (45)	Yes	Yes	Yes
Pseudo R2	0.114	0.134	0.138
Log likelihood	–2,255.5911	–2,204.253	–2,194.1294
Observations	*4,988*	*4,988*	*4,988*

Notes: Marginal effects displayed.
Standard errors in parentheses.
*Significant at 5%.
**Significant at 1%.
***Significant at 0.1%.

4 Self-selection of workers with regard to difficulty

Apart from their direct execution, the management of difficult tasks in microwork environments is also affected by workers' choice about which tasks to execute since, as in all crowd sourced environments, the global allocation of efforts results from the aggregation of workers' individual choices among available problems.[13] Recent evidence from large open-source software projects has indeed suggested a relationship between task difficulty and the allocation of efforts: teamwork is found more frequently on files containing more complex software procedures (den Besten et al., 2008). Compared to other communities such as open-source software or Wikipedia, where coordination and semantic signalling seems prevalent (Rossi et al., 2010; den Besten and Dalle, 2014), individual choices in AMT are likely to be affected by price signals in addition to other characteristics of tasks and projects.

To investigate the determinants of workers' choice in AMT, we built a dataset by crawling AMT's website every three minutes during a period of approximately two months, enabling us to gather data on several hundreds of real AMT projects (Dalle et al., 2014). The supply of labour is observed indirectly, by measuring the mean speed at which individual tasks are executed from available projects. Table 2 presents the results of OLS regressions with this dependent variable (mean speed of execution) and using the price of individual tasks in the project, the maximum number of tasks in the project or its natural logarithm, the length of the title given to the project or the length of the description given for individual tasks, as regressors. To improve the robustness of our results, we limited our sample to projects for which we had sufficiently consistent data and to projects for which the 'master' qualification granted by Amazon to its regular workers had been requested (see Section 2).

Table 2 OLS regression

Dependent variable: mean speed of execution of individual tasks within projects				
	(1)	*(2)*	*(3)*	*(4)*
Price (monetary reward per individual task)	–1.010 (0.343)	–0.136 (1.556)	–1.659 (1.420)	0.781 (1.447)
Size (maximum number of individual tasks in the project during the period studied)	0.0000872*** (0.0000189)		0.0000819*** (0.0000177)	
Log (Size) (logarithm of the former)		1.249*** (0.280)		1.190*** (0.262)
Length (Title) (number of characters in the title given by the requester to the project)	–0.00773 (0.00688)	–0.00613 (0.00690)		
Length (Desc) (number of characters in the description given by the requester to the tasks)			0.00531*** (0.00143)	0.00532*** (0.00144)
Intercept	0.730* (0.344)	–3.131** (0.971)	–0.0878 (0.216)	–3.706*** (0.876)
Adjusted R2	0.18	0.17	0.28	0.28
F-statistic (p-value)	0.000145	0.000244	<0.0001	<0.0001
Observations (projects)	*91*	*91*	*91*	*91*

Notes: Standard errors in parentheses.
*Significant at 5%.
**Significant at 1%.
***Significant at 0.1%.

The price of individual tasks is not significant in any of the regressions, a result that could seem counter-intuitive but is easily explained if online microworkers maximise their income (Dalle et al., 2014): income being given not simply by the price of individual tasks, but also by the efficiency in executing them correctly, a characteristic

for which their difficulty is obviously relevant. In the previous experiment (Section 3), giving correct answers for author pairs looking as positive matches (high global score) which are in fact bad ones (not validated by experts) would have been feasible but would have required more time and/or efforts from workers: the screen seen by workers in our AMT experiment listed the names and affiliations of the authors and inventors to be matched, as well as the titles of their corresponding papers and patents hyperlinked to the documents available online. Furthermore, workers, notably at the master level, can have incentives to work on easily doable tasks instead of more difficult ones, since the acceptance rate of their work, based on their performance, would be improved with this strategy – and is key to gain and retain the master level.

Our results are consistent with these explanations since

a the length of the description of these tasks (a proxy for the quality of the design and the detailed description of these tasks)is positively correlated to the supply of online labour, contrary to the length of the title, a similar variable used in order to rule out other explanations

b the size of the project is very significantly correlated to workers' choice, which would be expected if workers tend to specialise to improve their productivity and their income.

These observations are also coherent with Franklin et al. (2011).

We then re-posted the same tasks descriptions we had in the previous database on AMT, asking online workers how they evaluated their difficulty (from very easy to very difficult on a 5-level scale): first without disclosing the price of individual tasks, and second while including it in the evaluation. We observed that workers appeared sensitive to price signals in the context of their assessment of the difficulty of the tasks (see Figures 2 and 3). This observation provides a possible further explanation for the non-significance of price in the former regressions, that is consistent not only with Yan et al. (2010)'s analysis according to which low-priced tasks tend to be addressed more rapidly, but also with earlier work on the assessment of difficulty which stressed the existence of both an objective and a subjective component (Maynard and Hakel, 1997). Microworkers would react to price signals but only by subjectively inferring the difficulty of the tasks proposed from their price.

Finally, these observations suggest an alternative explanation for the results of Mason and Watts (2010), who showed that performance depended mostly on motivation and much less on pay for certain types of tasks (sorting images) at least, while observing further that there was 'no interaction between difficulty and compensation'. The fact that the price of individual tasks did not affect quality in their setting, and was independent from difficulty, might indeed result from the self selection of workers and Mason and Watts do mention sorting effects. Price signals are interpreted by workers in the context of their wage expectations that depend on the difficulties of the tasks proposed. Mason and Watts were able to attract more motivated workers when they proposed more interesting tasks (crossword puzzles); conversely, when simpler tasks of varying but implicit difficult levels were proposed, they might simply have been executed by workers minimising their opportunity costs of not working on simple tasks, i.e., executing them as a routine – as 'no-brainers', so to say.

Figure 2 Workers evaluate difficulty without any information on pricing

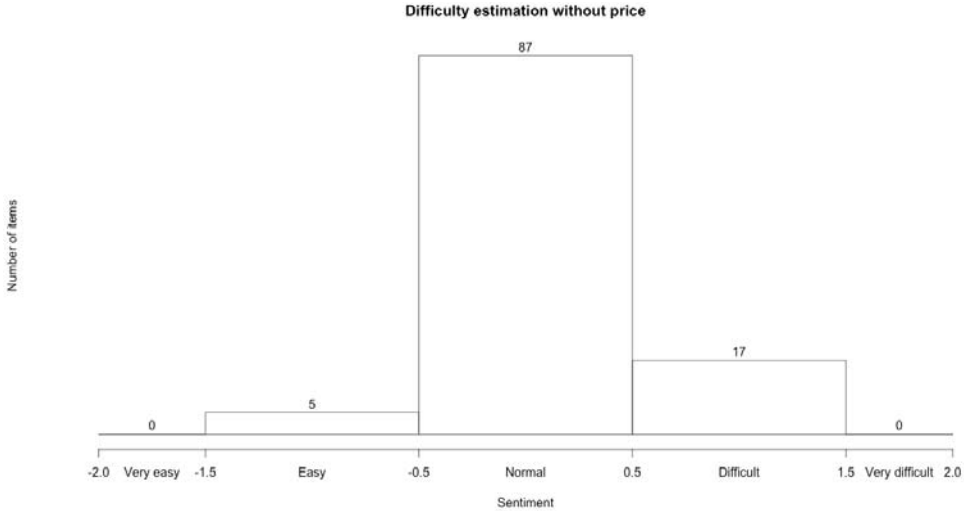
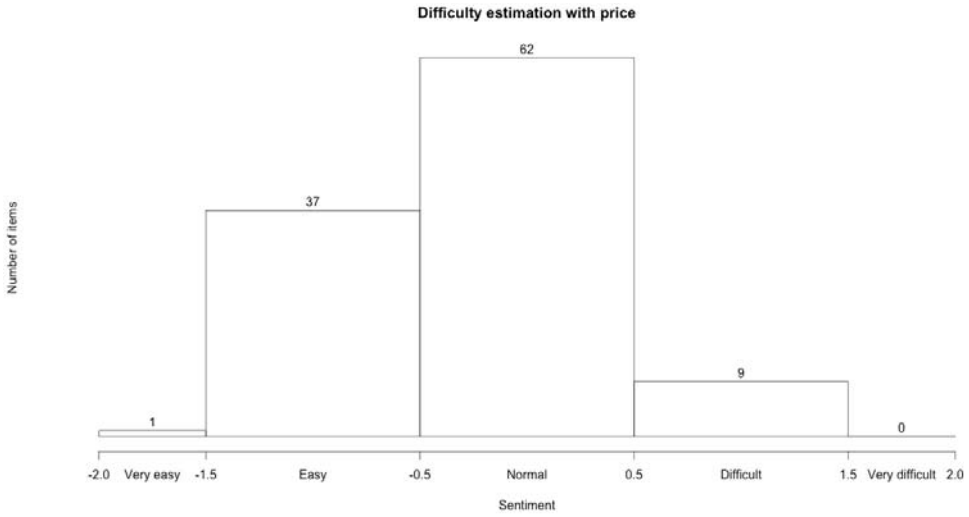


Figure 3 Workers are given price information when evaluating difficulty



5 Discussion and conclusions

In this article, we have identified three problems associated with the management of difficult tasks in microwork environments: first, difficulty is generally unknown ex ante, since tasks of a similar nature but of varying difficulties are mixed without their difficulty being apparent; second, unskilled problem solvers can fail without noticing their errors, thus ruling out solutions that would use failure signals to design upgraded crowdsourcing mechanisms; and last, but not least, the platform dimension of microwork implies that workers can self-select against difficult tasks to improve their income and their

track-record. This last issue is even more acute as price signals might actually be interpreted by workers in the context of evaluating the difficulty of tasks[14], at least on AMT.

However, the specific design of AMT as a microwork platform may simply not be an appropriate tool for all kinds of tasks and notably for the most difficult ones. Indeed, in their detailed survey of seven online platforms beyond AMT, most of which qualify as microwork platforms, Vakharia and Lease (2013) mention that several platforms attempt to provide 'specialised and complex task support': for instance, ClickWorker has a special focus on text building tasks for search engine optimisation, CrowdSource on writing and text creation tasks, MobileWorks on testing iOS apps and games, etc.

Platform specialisation, generally associated with the training of workers, actually appears as a possible platform-based remedy to the problems associated with the management of difficult tasks. Platform innovations such as the recruitment of workers tied to the platform; or the internal use of the platform where requesters offer tasks only to their own private work force; or else mechanisms proposed to route tasks to the most appropriate workers, can also be interpreted in this sense. In addition, most of these platforms offer schemes to promote worker participation and allow for the modulation of rewards based on characteristics of the tasks, while offering mechanisms to associate reputation and skills to online workers[15].

We suggest that through their effort to monitor and develop their workforce, these microwork platforms make a differentiating claim, even implicit, that they can overcome the problems typically faced by AMT when handling difficult tasks. Also, significantly, oDesk, a platform that would probably not qualify as a microwork platform as it is project-based, allows workers based in US or Canada to be entitled to employee benefits like health insurance or 401(k) retirement saving plans. In this case, it is the contract proposed to workers that is key to select them, as in a standard separating contract framework: an observation also coherent with classical labour economics suggesting explanations for the emergence of permanent contract in agrarian economies due to the existence of "important tasks that require judgment, discretion and care" (Eswaran and Kotwal, 1985).

These findings and observations, and the evolution of microwork platforms, are perfectly in line with the recent results of Staffelbach et al. (2015), who compare results from AMT workers with those of civil engineering graduate students in the context of the analysis of wind simulation data. They conclude that increased communication was needed for the handling of more complex tasks, and that the employment relationship then began to be closer to outsourcing. As they put it, "rather than allowing crowd workers to replace expert labor, which would be dangerous, we envision the role of crowd workers as an assistant to expert labor, since the time and effort of expert labor is limited and expensive".

We suggest that "identifying which tasks to farm out and who within your organization should manage them" (Boudreau and Lakhani, 2013) is indeed a problematic issue for the 'spot' online labour markets whose existence is permitted by specialised microwork platforms. Experts might be needed to decide how labour should be divided between the firm and the crowd, i.e., notably between difficult and less difficult tasks, or at least to monitor, detect and notice errors made by less skilled workers in order to render mechanisms à la Acemoglu et al. (2014) implementable. But the demand for better contractual solutions by these experts should probably impact these evolutions and might transform specialised microwork platforms into quasi-outsourcing

companies with a salaried expert workforce. The price of microwork would certainly increase, as a consequence, since the costs of managing difficult tasks would thus be internalised. Since part of the promise of microwork is to allow for the treatment of massive and heterogeneous data at a reasonable cost, future studies should then certainly inquire about the elasticity of the demand of microwork to its price.

Acknowledgements

We would like to thank Nicolas Maisonneuve, Vili Lehdonvirta, Francesco Lissoni and Pluvia Zuniga. We acknowledge funding from the European Science Foundation Research Networking Programme Academic Patenting in Europe (APE-INV), the 2012 short-term visiting fellows program of the CSIC Institute of Public Goods and Policies (CSIC-IPP), the Spanish Ministry of Economics and Competitiveness (CSO2009-10845 and CSO2012-32844) and the Agence Nationale de Recherche (ANR) for project 'OCKTOPUS' (ANR¬12¬CORD¬026), as well as through the program 'Investments for the Future' and the chair 'Entrepreneurship & Innovation' (ANR-10-LabX-11-01). We would also like to thank the editor and three anonymous referees for their very helpful suggestions on a previous version.

References

Acemoglu, D., Mostagir, M. and Ozdaglar, A. (2014) *Managing Innovation in a Crowd*, No. w19852, National Bureau of Economic Research.

Ahmad, S., Battle, A., Malkani, Z. and Kamvar, S. (2011) 'The jabberwocky programming environment for structured social computing', *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, October, pp.53–64, ACM.

Alonso, O. and Mizzaro, S. (2012) 'Using crowdsourcing for TREC relevance assessment', *Information Processing & Management*, Vol. 48, No. 6, pp.1053–1066.

Bergvall-Kåreborn, B. and Howcroft, D. (2014) 'Amazon Mechanical Turk and the commodification of labour', *New Technology, Work and Employment*, Vol. 29, No. 3, pp.213–223.

Bernstein, M.S., Little, G., Miller, R.C., Hartmann, B., Ackerman, M.S., Karger, D., Crowell, D. and Panovich, K. (2010) 'Soylent: a word processor with a crowd inside', *Proceedings of the 23rd Annual ACM Symposium on User Interface Software and Technology, UIST'10*, pp.313–322, ACM, New York, NY, USA, 2010. doi:10.1145/1866029.1866078.

Boudreau, K. (2010) 'Open platform strategies and innovation: granting access vs. devolving control', *Management Science*, Vol. 56, No. 10, pp.1849–1872.

Boudreau, K.J. and Lakhani, K.R. (2013) 'Using the crowd as an innovation partner', *Harvard Business Review*, Vol. 91, No. 4, pp.60–69.

Carayol, N. and Dalle, J.M. (2007) 'Sequential problem choice and the reward system in open science', *Structural Change and Economic Dynamics*, Vol. 18, No. 2, pp.167–191.

Ceccagnoli, M., Forman, C., Huang, P. and Wu, D.J. (2012) 'Co-creation of value in a platform ecosystem: the case of enterprise software', *MIS Quarterly*, Vol. 36, No. 1, pp.263–290.

Chandler, J., Mueller, P. and Paolacci, G. (2014) 'Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers', *Behavior Research Methods*, Vol. 46, No. 1, pp.112–130, DOI 10.3758/s13428-013-0365-7.

Curchod, C. and Neysen, N. (2009) *Disentangling Positive and Negative Externalities on Two-Sided Markets: The Ebay Case* [online] http://www.uclouvain.be/cps/ucl/doc/iag/documents/WP_3_Curchod_Neysen.pdf (accessed 21 September 2016).

Cuxac, P., Lamirel, J.C. and Bonvallot, V. (2012) 'Efficient supervised and semi-supervised approaches for affiliations disambiguation', *Scientometrics*, Vol. 97, No. 1, pp.1–12.

Dalle, J.M. and David, P.A. (2005) 'The allocation of software development resources in 'open source'', *Perspectives on Open Source and Free Software*, MIT Press, Cambridge, MA [online] http://siepr.stanford.edu/papers/pdf/02-27.pdf (accessed 21 September 2016).

Dalle, J.M., Lacroix, T., Lacage, M. and den Besten, M. (2014) 'A direct empirical study of the determinants of online work supply in Amazon Mechanical Turk', *Proceedings of IPP2014: Crowdsourcing for Politics and Policy*, September, Oxford.

Demartini, G., Difallah, D.E. and Cudré-Mauroux, P. (2012) 'ZenCrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking', *Proceedings of the 21st International Conference on World Wide Web, WWW'12*, pp.469–478, ACM, New York, NY, USA.

den Besten, M. and Dalle, J.M. (2014) 'Coordination by reassignment in the Firefox community', *Proceedings of ECIS 2014*, July, Tel Aviv.

den Besten, M., Dalle, J.M. and Galia, F. (2008) 'The allocation of collaborative efforts in open-source software', *Information Economics and Policy*, Vol. 20, No. 4, pp.316–322.

Eickhoff, C. and de Vries, A.P. (2013) 'Increasing cheat robustness of crowdsourcing tasks', *Information Retrieval*, Vol. 16, No. 2, pp.121–137, doi:10.1007/s10791-011-9181-9.

Ester, M., Kriegel, H.P., Sander, J. and Xu, X. (1996) 'A density-based algorithm for discovering clusters in large spatial databases with noise', in Simoudis, E., Han, J. and Fayyad, U.M. (Ed.): *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pp.226–231, AAAI Press.

Eswaran, M. and Kotwal, A. (1985) 'A theory of two-tier labor markets in agrarian economies', *American Economic Review*, Vol. 75, No. 1, pp.162–177.

Faridani, S., Hartmann, B. and Ipeirotis, P.G. (2011) 'Whats the right price? Pricing tasks for finishing on time, human computation', *AAAI Workshop (WS-11-11)* [online] https://www.aaai.org/ocs/index.php/WS/AAAIW11/paper/viewFile/3994/4269 (accessed 21 September 2016).

Franklin, M., Kossman, D., Kraska, T., Ramesh, S. and Xin, R. (2011) 'CrowdDB: answering queries with crowdsourcing', *SIGMOD '11*, 12–16 June, Athens, Greece [online] https://people.eecs.berkeley.edu/~franklin/Papers/CrowdDBSigmod11.pdf (accessed 21 September 2016).

Gawer, A. (2014) 'Bridging differing perspectives on technological platforms: toward an integrative framework', *Research Policy*, Vol. 43, No. 7, pp.1239–1249.

Gottlieb, L., Friedland, G., Choi, J.Y., Kelm, P. and Sikora, T. (2014) 'Creating experts from the crowd: techniques for finding workers for difficult tasks', *IEEE Transactions on Multimedia*, Vol. 16, No. 7, pp.2075–2079.

Gurney, T., Horlings, E. and van den Besselaar, P. (2012) 'Author disambiguation using multi-aspect similarity indicators', *Scientometrics*, Vol. 91, No. 2, pp.435–449.

Hirth, M., Hoßfeld, T. and Tran-Gia, P. (2013) 'Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms', *Mathematical and Computer Modelling*, Vol. 57, No. 11, pp.2918–2932.

Hong, L. and Page, S.E. (2004) 'Groups of diverse problem solvers can outperform groups of high-ability problem solvers', *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 101, No. 46, pp.16385–16389.

Horton, J. and Chilton, L. (2010) *The Labor Economics of Paid Crowdsourcing*, 1001.0627, 5 January [online] http://arxiv.org/abs/1001.0627 (accessed 21 September 2016).

Ipeirotis, P. (2010a) *Getting High Quality Results on MTurk | A Computer Scientist in a Business School* [online] http://www.behind-the-enemy-lines.com/2010/03/getting-high-quality-results-on-mturk.html (accessed 28 January 2013).

Ipeirotis, P. (2010b) *Be a Top Mechanical Turk Worker: You Need $5 and 5 Minutes | A Computer Scientist in a Business School* [online] http://www.behind-the-enemy-lines.com/2010/10/be-top-mechanical-turk-worker-you-need.html (accessed 28 January 2013).

Ipeirotis, P. (2012) *Mechanical Turk Changing the Defaults: The Game has Changed | A Computer Scientist in a Business School* [online] http://www.behind-the-enemy-lines.com/2012/12/mechanical-turk-changing-defaults-game.html (accessed 28 January 2013).

Irani, L. (2015) 'The cultural work of microwork', *New Media & Society*, Vol. 17, pp.720–739.

Kittur, A., Chi, E.H. and Suh, B. (2008) 'Crowdsourcing user studies with Mechanical Turk', *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, *CHI '08*, pp.453–456, ACM, New York, NY, USA, doi:10.1145/1357054.1357127.

Kittur, A., Nickerson, J., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M. and Horton, J. (2012) *The Future of Crowd Work*, SSRN working paper, 19 December [online] http://papers.ssrn.com/abstract=2190946 (accessed 21 September 2016).

Kittur, A., Smus, B., Khamkar, S. and Kraut, R.E. (2011) 'Crowdforge: crowdsourcing complex work', *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, October, pp.43–52, ACM.

Kulkarni, A., Can, M. and Hartmann, B. (2012) 'Collaboratively crowdsourcing workflows with turkomatic', *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, February, pp.1003–1012, ACM.

Lauritzen, G.D., Salomo, S. and La Cour, A. (2013) 'Dynamic boundaries of user communities: exploiting synergies rather than managing dilemmas', *International Journal of Technology Management*, Vol. 63, No. 3, pp.148–168.

Lehdonvirta, V. (2016) 'Algorithms that divide and unite: delocalization, identity, and collective action in 'microwork'', to appear in Flecker, J. (Ed.): *Space, Place and Global Digital Work*, Palgrave-Macmillan, London.

Lewis, S.C., Zamith, R. and Hermida, A. (2013) 'Content analysis in an era of big data: a hybrid approach to computational and manual methods', *Journal of Broadcasting & Electronic Media*, Vol. 57, No. 1, pp.34–52.

Maisonneuve, N. and Chopard, B. (2012) 'Crowdsourcing satellite imagery analysis: study of parallel and iterative models', *Geographic Information Science*, pp.116–131, Springer, Berlin Heidelberg.

Maraut, S. and Martínez, C. (2014) 'Identifying author – inventors from Spain: methods and a first insight into results', *Scientometrics*, Vol. 101, No. 1, pp.445–476, doi: 10.1007/s11192-014-1409-1.

Marge, M., Banerjee, S. and Rudnicky, A.I. (2010) 'Using the Amazon Mechanical Turk to transcribe and annotate meeting speech for extractive summarization', *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, CSLDAMT '10*, pp.99–107, ACL, Stroudsburg, PA, USA [online] http://dl.acm.org/citation.cfm?id=1866696.1866712 (accessed 21 September 2016).

Mason, W. and Suri, S. (2012) 'Conducting behavioral research on Amazon's Mechanical Turk', *Behavior Research Methods*, Vol. 44, No. 1, pp.1–23, doi:10.3758/s13428-011-0124-6.

Mason, W. and Watts, D.J. (2010) 'Financial incentives and the performance of crowds', *ACM SigKDD Explorations Newsletter*, Vol. 11, No. 2, pp.100–108.

Maynard, D.C. and Hakel, M.D. (1997) 'Effects of objective and subjective task complexity on performance', *Human Performance*, Vol. 10, No. 4, pp.303–330.

Noronha, J., Hysen, E., Zhang, H. and Gajos, K.Z. (2011) 'Platemate: crowdsourcing nutritional analysis from food photographs', *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, October, pp.1–12, ACM.

Raasch, C. (2011) 'The sticks and carrots of integrating users into product development', *International Journal of Technology Management*, Vol. 56, No. 1, pp.21–39.

Rand, D.G. (2012) 'The promise of Mechanical Turk: how online labor markets can help theorists run behavioral experiments', *Journal of Theoretical Biology*, April, Vol. 299, pp.172–179, doi:10.1016/j.jtbi.2011.03.004.

Rong, K., Lin, Y., Shi, Y. and Yu, J. (2013) 'Linking business ecosystem lifecycle with platform strategy: a triple view of technology, application and organization', *International Journal of Technology Management*, Vol. 62, No. 1, pp.75–94.

Rossi, A., Gaio, L., den Besten, M. and Dalle, J.M. (2010) 'Coordination and division of labor in open content communities: the role of template messages in Wikipedia', *43rd Hawaii International Conference on System Sciences (HICSS 2010)*, January, pp.1–10, IEEE.

Shaw, A.D., Horton, J.J. and Chen, D.L. (2011) 'Designing incentives for inexpert human raters', *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work, CSCW '11*, pp.275–284, ACM, New York, NY, USA.

Smalheiser, N.R. and Torvik, V.I. (2009) 'Author name disambiguation', *Annual Review of Information Science and Technology*, Vol. 43, No. 1, pp.1–43.

Snow, R., O'Connor, B., Jurafsky, D. and Ng, A.Y. (2008) 'Cheap and fast – but is it good?: Evaluating non-expert annotations for natural language tasks', *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08*, pp.254–263, ACL, Stroudsburg, PA, USA [online] http://dl.acm.org/citation.cfm?id=1613715.1613751 (accessed 21 September 2016).

Staffelbach, M., Sempolinski, P., Kijewski-Correa, T., Thain, D., Wei, D., Kareem, A. and Madey, G. (2015) 'Lessons learned from crowdsourcing complex engineering tasks', *PloS one*, Vol. 10, No. 9, p.e0134978.

Rayna, T. and Striukova, L. (2015) 'Open innovation 2.0: is co-creation the ultimate challenge?', *International Journal of Technology Management*, Vol. 69, No. 1, pp.38–53.

Thomas, L.D., Autio, E. and Gann, D.M. (2014) 'Architectural leverage: putting platforms in context', *The Academy of Management Perspectives*, Vol. 28, No. 2, pp.198–219.

Toomim, M., Kriplean, T., Pörtner, C. and Landay, J. (2011) 'Utility of human-computer interactions: toward a science of preference measurement', in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, May, pp.2275–2284, ACM.

Vakharia, D. and Lease, M. (2013) *Beyond AMT: An Analysis of Crowd Work Platforms*, arXiv e-print No. 1310.1672 [online] http://arxiv.org/abs/1310.1672 (accessed 21 September 2016).

Veve, M. (2009) 'Supporting name authority control in XML metadata: a practical approach at the University of Tennessee', *Library Resources & Technical Services*, Vol. 53, No. 1, pp.41–52.

von Ahn, L., Blum, M., Hopper, N.J. and Langford, J. (2003) 'CAPTCHA: using hard AI problems for security', *Advances in Cryptology – EUROCRYPT 2003*, pp.294–311, Springer, Berlin Heidelberg.

von Ahn, L., Maurer, B., McMillen, C., Abraham, D. and Blum, M. (2008) 'Recaptcha: human-based character recognition via web security measures', *Science*, Vol. 321, No. 5895, pp.1465–1468.

Wang, J., Berzins, K., Hicks, D., Melkers, J., Xiao, F. and Pinheiro, D. (2012a) 'A boosted-trees method for name disambiguation', *Scientometrics*, Vol. 93, No. 2, pp.391–411.

Wang, J., Kraska, T., Franklin, M.J. and Feng, J. (2012b) 'CrowdER: crowdsourcing entity resolution', *Proc. VLDB Endow.*, Vol. 5, No. 11, pp.1483–1494.

Yan, T., Kumar, V. and Ganesan, D. (2010) 'Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones', *Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services*, June, pp.77–90, ACM.

Notes

1 https://en.wikipedia.org/wiki/Human-based_computation.

2 On related issues associated with crowdsourcing and with user- and open-innovation with respect to platforms, see, e.g., Raasch (2011), Lauritzen et al. (2013) and Rayna and Striukova (2015).

3 Compared to platforms such as, e.g., TaskRabbit that deal with material tasks.

4 This focus needs to be challenged and might even have blurred some of the relevant issues: see Section 5.

5 For information on most recent developments about the so-called 'Names Game' in patent data see the European Science Foundation Research Networking Programme – Academic Patenting in Europe (APE-INV) at http://www.esf-ape-inv.eu/.

6 In the a first phase, candidate author and inventor names likely to belong to the same person were automatically grouped together in clusters and linked back to their corresponding patent applications and publications. Clusters were built based on name matching and disambiguation variables calculated using all available information in the papers and patents from the authors and inventors, such as name, institutional affiliation, discipline, geographical location, etc. The technique to construct them was DBSCAN (Ester et al., 1996), a density-based technique which relies on the notion of density reachability and connectivity. In a second phase, expert reviewers checked manually a number of dubious matches flagged automatically in the first phase. Thanks to this manual revision, false matches were excluded and only validated matches were used to improve the disambiguation recursively and improve precision of the final dataset. Dubious matches would typically be patent-paper pairs whose validity is difficult to assess due to common names, spelling mistakes, mobility (different affiliations) or multi-disciplinarity (different areas of specialisation) of their corresponding authors and inventors.

7 Clusters including pairs with uncommon names of authors and inventors that were exactly matched (e.g., no spelling mistakes) were deliberately excluded from this sample in order to avoid offering too simple tasks to AMT workers.

8 We set the price at USD 0.20 based on findings from previous studies (e.g., Faridani et al., 2011; Horton and Chilton, 2010). In addition, before launching the experiment itself, we submitted several batches of ten tasks each, each of which could be resolved by a maximum of five workers with a prospective reward per task of USD 0.05 and USD 0.10 respectively, limiting their visibility to workers with a good record. However, these batches failed to elicit sufficient response: in each, only three tasks were completed over the next days by one worker in the first batch and three distinct workers respectively in the second batch. We then submitted two more batches offering a much higher reward of USD 0.50 per task completed. We made the first batch visible to all workers on the platform; the second only to those for whom more than 60% of past work had been approved by requesters. This time, all tasks were completed within a day after publication by 17 workers. Based on that, we decided to choose an intermediate price for the real experiment, at USD 0.20 per HIT.

9 The Gini coefficient for the number of tasks submitted per worker is 0.75.

10 A Pearson product-moment correlation between launch dates of batches and their completion rate is equal to 0.7, significant at 5%.

11 The two batches launched in the weekend have of 0.4 and 0.18 compare to the median and mean of 0.6 for batches launched during weekdays.

12 Maraut and Martínez (2014) calculated the global score for about half of the possible combination of pairs in the clusters and propagated the validation to the full sample of 11,918 by applying transitivity. For the statistical analysis presented here, we only use the pairs with a global score assigned, before applying transitivity. This explains the smaller sample size in the logit regression presented in Table 1.

13 A general issue known as the ‘problem of problem choice’ (Dalle and David, 2005; Carayol and Dalle, 2007).

14 Further studies could inquire whether there would be some connection between the other professional occupations of microworkers, their marital status and their family situation, and the nature of the tasks they tend to address.

15 These evolutions contrast significantly with recent developments on AMT that have mostly focused on adverse selection issues, with Turk-Opticon for workers and the introduction of qualifications for requesters.