Clarity is a Worthwhile Quality – On the Role of Task Clarity in Microtask Crowdsourcing

Ujwal Gadiraju
L3S Research Center
Leibniz Universität Hannover
Hannover, Germany
gadiraju@L3S.de

Jie Yang
Web Information Systems
Delft University of Technology
The Netherlands
j.yang-3@tudelft.nl

Alessandro Bozzon
Web Information Systems
Delft University of Technology
The Netherlands
a.bozzon@tudelft.nl

ABSTRACT

Workers of microtask crowdsourcing marketplaces strive to find a balance between the need for monetary income and the need for high reputation. Such balance is often threatened by poorly formulated tasks, as workers attempt their execution despite a sub-optimal understanding of the work to be done.

In this paper we highlight the role of *clarity* as a characterising property of tasks in crowdsourcing. We surveyed 100 workers of the CrowdFlower platform to verify the presence of issues with task clarity in crowdsourcing marketplaces, reveal how crowd workers deal with such issues, and motivate the need for mechanisms that can predict and measure task clarity. Next, we propose a novel model for task clarity based on the *goal* and *role* clarity constructs. We sampled 7.1K tasks from the Amazon mTurk marketplace, and acquired labels for task clarity from crowd workers. We show that task clarity is coherently perceived by crowd workers, and is affected by the type of the task. We then propose a set of features to capture task clarity, and use the acquired labels to train and validate a supervised machine learning model for task clarity prediction. Finally, we perform a long-term analysis of the evolution of task clarity on Amazon mTurk, and show that clarity is not a property suitable for temporal characterisation.

CCS CONCEPTS

•Information systems → World Wide Web; Crowdsourcing;

KEYWORDS

Task Clarity; Crowdsourcing; Microtasks; Crowd Workers; Goal Clarity; Role Clarity; Prediction; Performance

1 INTRODUCTION

Microtask crowdsourcing has become an appealing approach for data collection and augmentation purposes, as demonstrated by the consistent growth of crowdsourcing marketplaces such as Amazon Mechanical Turk[1] and CrowdFlower[2].

[1]http://www/mturk.com/
[2]http://www.crowdflower.com/

HT'17, July 4-7, 2017, Prague, Czech Republic.

DOI: http://dx.doi.org/10.1145/3078714.3078715

Task consumption in microtask crowdsourcing platforms is mostly driven by a self-selection process, where workers meeting the required eligibility criteria select the tasks that they prefer to work on. Workers strive to maintain high reputation and performance to access more tasks, while maximizing monetary income. When discussing such a trade-off, the dominant narrative suggests that workers are more interested in obtaining their rewards, than in executing good work. We challenge this widespread opinion by focusing on an often neglected component of microtask crowdsourcing: the *clarity* of task description and instructions in terms of comprehensibility for workers.

Poor formulation of tasks has clear consequences: to compensate for a lack of alternatives in the marketplace, workers often attempt the execution of tasks despite a sub-optimal understanding of the work to be done. On the other hand, requesters are often not aware of issues with their task design, thus considering unsatisfactory work as evidence of malicious behaviour and deny rewards. As a result, crowd workers get demotivated, the overall quality of work produced decreases, and all actors lose confidence in the marketplace. Despite the intuitive importance of task *clarity* for microtask crowdsourcing, there is no clear understanding of the extent by which the lack of clarity in task description and instructions impacts worker performance, ultimately affecting the quality of work.

Research Questions and Original Contributions. This paper aims at filling this knowledge gap by contributing novel insights on the nature and importance of task clarity in microtask crowdsourcing. By combining qualitative and quantitative analysis, we seek to answer the following research questions.

> **RQ1**: What makes the specification of a task unclear to crowd workers? How do workers deal with unclear tasks?

First, we investigate if clarity is indeed a concern for workers. We designed and deployed a survey on the CrowdFlower platform, where we asked workers to describe what makes a task unclear, and to illustrate their strategies for dealing with unclear tasks. The survey involved 100 workers, and clearly highlights that workers confront unclear tasks on a regular basis.

Some workers attempt to overcome the difficulties they face with inadequate instructions, and unclear language by using external help, dictionaries or translators. Several workers tend to complete unclear tasks despite not understanding the objectives entirely.

These results demonstrate the need for methods for task clarity measurement and prediction, and shaped the formulation of the following questions.

RQ2: How is the clarity of crowdsourcing tasks perceived by workers, and distributed over tasks?

Inspired by work performed in the field of organisational psychology, we consider clarity both in the context of *what* needs to be produced by the worker (*goal clarity*) and *how* such work should be performed (*role clarity*). We sampled 7.1K tasks from a 5 years worth dataset of the Amazon mTurk marketplace. Tasks were published on CrowdFlower to collect clarity assessments from workers. Results show that task clarity is coherently perceived by crowd workers, and is affected by the type of the task. We unveil a significant lack of correlation between the *clarity* and the *complexity* of tasks, thus showing that these two properties orthogonally characterise microwork tasks.

RQ3: Which features can characterise the goal and role clarity of a task? Using such features, to what extent can task clarity be predicted?

We propose a set of features based on the metadata of tasks, task type, task content, and task readability to capture task clarity. We use the acquired labels to train and validate a supervised machine learning model for task clarity prediction. Our proposed model to predict task clarity on a 5-point scale achieves a mean absolute error (*MAE*) of *0.4 (SD=.003)*, indicating that task clarity can be accurately predicted.

RQ4: To what extent is task clarity a macro-property of the Amazon mTurk ecosystem?

We analyzed 7.1K tasks to understand how task clarity evolves over time. We found that the overall task clarity in the marketplace fluctuates over time, albeit without a discernible pattern. We found a weak positive correlation between the average task clarity and the number of tasks deployed by requesters over time, but no significant effect of the number of tasks deployed by requesters on the magnitude of change in task clarity.

2 RELATED LITERATURE

Text readability. Readability has been defined as the sum of all elements in text that affect a reader's understanding, reading speed and level of interest in the material [11]. There has been a lot of work in the past on analyzing the readability of text, as summarized in [7]. Early works range from simple approaches that focus on the semantic and syntactic complexity of text [25], or vocabulary based approaches where semantic difficulty is operationalized by means of gathering information on the average vocabulary of a certain age or social status group [6]. More recently, authors proposed statistical language models to compute readability [8]. Other works studied the lexical richness of text by capturing the range and diversity of vocabulary in given text [29]. Several machine learning models have also been proposed to predict the readability of text [23, 31]. De Clerq et al. recently investigated the use of crowdsourcing for assessing readability [12]. The vast body of literature corresponding to text readability has also resulted in several software packages and tools to compute readability [9, 17].

In this paper, we draw inspiration from related literature on text readability in order to construct features that aid in the prediction of task clarity on crowdsourcing platforms.

Task Clarity in Microtask Crowdsourcing. Research works in the field of microtask crowdsourcing have referred to the importance of task clarity tangentially; several authors have stressed about the positive impact of task design, clear instructions and descriptions on the quality of crowdsourced work [3, 26, 30, 34]. Grady and Lease pointed out the importance of wording and terminology in designing crowdsourcing tasks effectively [16]. Alonso and Baeza-Yates recommended providing 'clear and colloquial' instructions as an important part of task design [1]. Kittur et al. identified 'improving task design through better communication' as one of the pivotal next steps in designing efficient crowdsourcing solutions in the future [27]. The authors elaborated that task instructions are often ambiguous and incomplete, do not address boundary cases, and do not provide adequate examples. Khanna et al. studied usability barriers that were prevalent on Amazon mTurk (AMT), which prevented workers with little digital literacy skills from participating and completing work on AMT [24]. The authors showed that the task instructions, user interface, and the workers' cultural context corresponded to key usability barriers. To overcome such usability obstacles on AMT and better enable access and participation of low-income workers in India, the authors proposed the use of simplified user interfaces, simplified task instructions, and language localization. More recently, Yang et al. investigated the role of *task complexity* in worker performance, with an aim to better the understanding of task-related elements that aid or deter crowd work [37].

While the importance of task clarity has been acknowledged in the microtask crowdsourcing community, there is neither a model that describes task clarity nor a measure to quantify it. In this paper, we not only propose a model for task clarity, but we also present a means to measure it. To the best of our knowledge, this is the first work that thoroughly investigates the features that determine task clarity in microtask crowdsourcing, and provides an analysis of the evolution of task clarity.

Task Clarity in Other Domains. In the field of organizational psychology, researchers have studied how the sexual composition of groups affects the authority behavior of group leaders in cases where the task clarity is either *high* or *low* [33]. In this case, the authors defined task clarity as the degree to which the goal (i.e., the desired outcome of an activity) and the role (i.e., the activities performed by an actor during the course of a task) are clear to a group leader. In self-regulated learning, researchers have widely studied task interpretation as summarized in [32]. Hadwin proposed a model that suggests the role of the following three aspects in task interpretation and understanding; (i) implicit aspects, (ii) explicit aspects, and socio-contextual aspects [18, 19]. Recent literature regarding task interpretation in the learning field has revolved around text decoding, instructional practices or perceptions of tasks on the one hand [4, 22, 28], and socio-contextual aspects

of task interpretation such as beliefs about expertise, ability, and knowledge on the other hand [5, 10].

Inspired by the modeling of task clarity in the context of authority behavior in Psychology, we model task clarity as a combination of *goal clarity* and *role clarity* (as explained in Section 4).

3 ARE CROWDSOUCRED MICROTASKS ALWAYS CLEAR?

We aim to investigate whether or not workers believe task clarity to impact their work performance (**RQ1**). We thereby deployed a survey consisting of various questions ranging from general demographics of the crowd to questions regarding their experiences while completing microtasks on crowdsourcing platforms.

3.1 Methodology

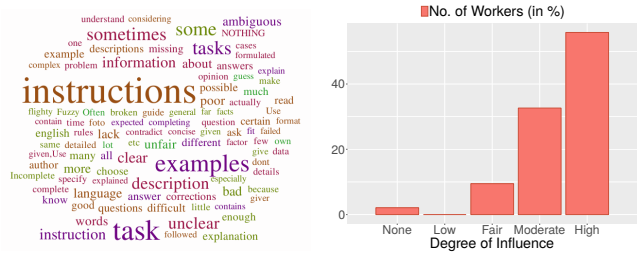
We deployed the survey on CrowdFlower[3] and gathered responses from 100 distinct crowd workers. To detect untrustworthy workers and ensure reliability of the responses received, we follow recommended guidelines for ensuring high quality results in surveys [15]. To this end, we intersperse two attention check questions within the survey. In addition, we use the filter provided by CrowdFlower to ensure the participation of only high quality workers (i.e., *level 3* crowd workers as prescribed on the CrowdFlower platform). We flagged workers who failed to pass at least one of the two attention check questions and do not consider them in our analysis.

3.2 Analysis and Findings

Worker's Experience. We found that around 36% of the workers who completed the survey earn their primary source of income through crowd work. 32.6% of the workers claim to have been contributing piecework through crowdsourcing platforms over the last 3 to 6 months. 63.2% of the workers have been doing so for the last 1 to 3 years. A small fraction of workers (3.2%) claim to have been working on microtasks for the last 3 to 5 years, while 1% of the worker population has been contributing to crowdsourced microtasks for over 5 years. During the course of this time, almost 74% of workers claim to have completed over 500 different tasks.

What factors make tasks unclear? We asked the workers to provide details regarding the factors that they believe make tasks unclear, in an open text field. The word-cloud in Figure 1a represents the responses collected from the crowd workers. Workers complained about the task instructions and descriptions being '*vague*', '*blank*', '*unclear*', '*inconsistent*', '*imprecise*', '*ambiguous*', or '*poor*'. Workers also complained about the language used; '*too many words*', '*high standard of English*', '*broken English*', '*spelling*', and so forth. Workers also pointed out that adequate examples are seldom provided by requesters. Excerpts of these responses are presented on the companion webpage[4].

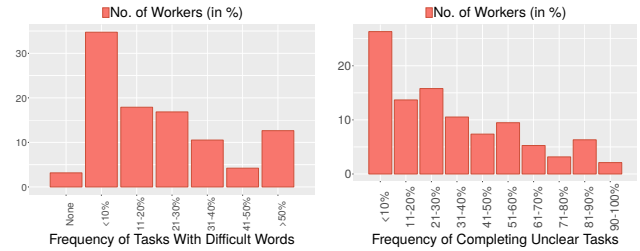
Task Clarity and Influence on Performance. Around 49% of workers claimed that up to a maximum of 30% of the tasks that they worked on were unclear. 37% of workers claimed that between 31-60% of the tasks they completed lacked clarity, while 14% of the workers claimed that more than 60% of their completed tasks were unclear. We also asked the workers about the perceived influence of



(a) (b)

Figure 1: (a) Word-cloud representing factors cited by workers that make tasks unclear. Size of words indicate their frequency. (b) Perceived degree of influence of task clarity on performance of workers.

task clarity on their performance. Our findings are presented in the Figure 1b. A large majority of workers believe that task clarity has a quantifiable influence on their performance. We also asked workers about the frequency of encounter for tasks containing difficult words, which might have hindered their performance. Figure 2a depicts our findings, indicating that workers observed tasks which contained difficult words reasonably frequently.



(a) (b)

Figure 2: (a) Frequency of tasks with difficult words, and (b) frequency of workers completing unclear tasks.

How do workers deal with unclear tasks? We investigated the frequency with which workers complete tasks despite the lack of clarity. As shown in Figure 2b, we found that nearly 27% of workers complete less than 10% of the unclear tasks that they encounter.

On the other hand, another 27% of workers completed more than 50% of all the unclear tasks they come across. In addition, around 18% of workers used dictionaries or other helpful means/tools to better understand over 50% of tasks they completed. 20% of workers used translators in more than 50% of the tasks that they completed.

4 MODELING TASK CLARITY

We address **RQ2** by modelling task clarity of crowdsourced microtasks as a combination of *goal clarity* and *role clarity*. Inspired by previous work in organizational psychology [33], we define task clarity as a combination of the extent to which the desired outcome of a task is clear (goal clarity), and the extent to which the workflow or activities to be carried out are clear (role clarity).

[3] http://crowdflower.com
[4] https://sites.google.com/site/ht2017clarity/

4.1 Assessing Task Clarity

Task clarity of microtasks in a marketplace is a notion that can be quantified by human assessors by examining task metadata such as the *title*, *keywords* associated with the task, *instructions* and *description*. Since these are the main attributes that requesters use to communicate the desired outcomes of the tasks, and prescribe how crowd workers should proceed in order to realize the objectives, we argue that they play an important role in shaping crowd work.

4.2 Acquiring Task Clarity Labels

With an aim to understand the distribution of task clarity across the diverse landscape of tasks on AMT [13], we sampled 7,100 tasks that were deployed on AMT over a period of 1 year between October 2013 to September 2014. For every month spanning the year, we randomly sampled 100 tasks of each of the 6 task types proposed in previous work [14]; *content creation* (CC), *information finding* (IF), *interpretation and analysis* (IA), *verification and validation* (VV), *content access* (CA)[5] and *surveys* (SU). Next, we deployed a job[6] on CrowdFlower to acquire task clarity labels from crowd workers. We first provided detailed instructions describing task clarity, goal clarity and role clarity. An excerpt from the task overview is presented below:

"Task clarity defines the quality of a task in terms of its comprehensibility. It is a combination of two aspects; (i) goal clarity, i.e, the extent to which the objective of the task is clear, and (ii) role clarity, i.e., the extent to which the steps or activities to be carried out in the task are clear."

In each task workers were required to answer 10 questions on a 5-point Likert scale. The questions involved assessing the goal and role clarity of the corresponding task, the overall task clarity, the influence of goal and role clarity in assessing overall task clarity, clarity of title, instructions and description, the extent to which the title conveyed the task description, the extent to which the keywords conveyed the task description and goal of the task, and the quality of language in the task description. Apart from these 10 questions, workers were provided with an optional text field where they could enter comments or remarks about the AMT task they evaluated. We gathered responses to these questions for each of the 7,100 tasks from 5 distinct crowd workers. We controlled for quality by using the *highest quality* restriction on CrowdFlower, that allows only workers with a near perfect accuracy over hundreds of different tasks and varying task types. In addition, we interspersed attention check questions where workers were asked to enter alphanumeric codes that were displayed to them. In return, workers were compensated according to the hourly rate of 7.5 USD.

4.3 Perception of Task Clarity

We found that the mean task, goal and role clarity across the different tasks were nearly the same. On average, workers perceived tasks to be moderately clear (*M=3.77, SD=.53*). The same is the case with goal clarity (*M=3.76, SD=.53*) and role clarity (*M=3.76, SD=0.54*). On investigating the influence of goal and role clarity on the crowd workers in adjudicating the overall task clarity, we found that role clarity and goal clarity were both important in determining task clarity. On average, workers responded that goal clarity influenced their judgment of overall task clarity to an extent of 3.98/5 (*SD=.51*), and that in case of role clarity was 3.93/5 (*SD=.52*). We found that goal clarity was slightly more influential than role clarity in determining the task clarity, and this difference was statistically significant; $t(14199) = 25.28, p < .001$.

We also analyzed the relationship of task clarity with goal and role clarity respectively. We found strong positive linear relationships in both cases, as shown in Figure 3.

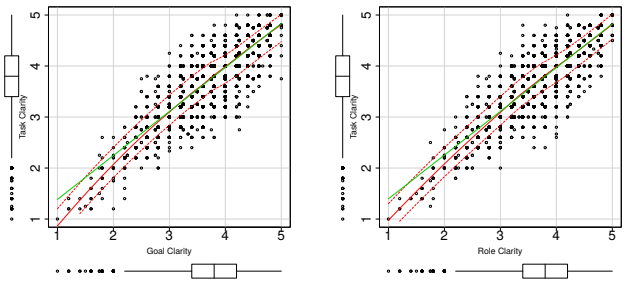


Figure 3: Relationship of Task Clarity with (a) Goal Clarity, and (b) Role Clarity. The trendline is represented in green, and the regression line is represented by the thick red line.

We computed Pearson's **r** between task clarity with each of goal and role clarity; $r(14998) = .85, R^2 = .72, p < .001$ and $r(14998) = .86, R^2 = .74, p < .001$. These findings indicate that it is equally important for task requesters to ensure that the objective of the task, as well as the means to achieve the desired outcome are adequately communicated to crowd workers via the task title, instructions and description, and keywords associated with the task.

4.3.1 Inter-worker Agreement. To find out whether or not task clarity is coherently perceived by workers, we verify the presence of agreement of task clarity evaluations among workers. Given the subjective nature of task clarity evaluations, we apply the SOS Hypothesis [20], which examines the extent to which individual evaluations of clarity spread around the mean clarity value per task. The SOS Hypothesis has proven to be more reliable than other inter-evaluator agreement measures such as Krippendorff's alpha, in subjective assessment tasks that involve a set of participants evaluating the same item – in our case, the same task [2]. In SOS Hypothesis, we test the magnitude of the squared relationship between the standard deviation (i.e. SOS) of the evaluations and the mean opinion score (MOS; in our case, mean clarity score), denoted by α. The value of α can then be compared with those of other subjective assessment tasks that are deemed to be more (high α) or less prone to disagreement (low α) among evaluators. Specifically, for 5-point scale evaluations, SOS Hypothesis tests a square relationship between SOS and MOS by fitting the following equation:

$$SOS(i) = -\alpha MOS(i)^2 + 6\alpha MOS(i) - 5\alpha$$

considering each task i in the evaluation pool.

[5]Note that there were fewer than 100 tasks of the CA type in a few months during the time period considered. In those cases, we considered all available tasks.
[6]Preview available in the companion page: https://sites.google.com/site/ht2017clarity/.

Table 1: SOS Hypothesis α values for Task Clarity, Goal Clarity and Role Clarity.

Clarity	Task Clarity	Goal Clarity	Role Clarity
α	0.3166	0.3229	0.3184

Table 1 shows the α values computed for task clarity, goal clarity and role clarity. All these evaluations have a value of 0.32, which is similar to what could be obtained in other subjective assessment tasks such as smoothness of web surfing, VoIP quality, and cloud gaming quality [20]. We therefore consider it acceptable. Figure 4 shows the fitted quadratic curve against worker evaluations for individual tasks. A significant correlation could be obtained between the fitted SOS value and the actual SOS value (Pearson's $r = 0.506$, $p < .001$). In conclusion, we find that task clarity is coherently perceived by workers. The substantial evidence of workers' agreement in perceiving task clarity helps establish the mean clarity score as ground truth for modeling task clarity using objective task features, as we report in Section 5.

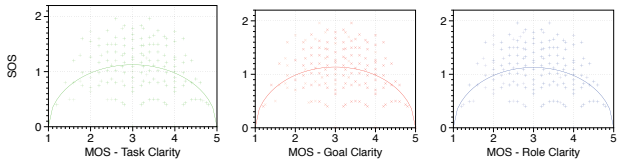


Figure 4: SOS Hypothesis plots for Task Clarity (green), Goal Clarity (red), and Role Clarity (blue). The quadratic curve depicts the fitting to worker evaluations for individual tasks.

4.3.2 Task Types and Perception of Task Clarity. We investigated the impact of task types on the perception of task clarity and the constructs of goal and role clarity. We note that Levene's test for homogeneity of variances was not violated across the different task types with respect to each of task, goal and role clarity. We conducted a one-way between workers ANOVAs to compare the effect of task types on the perception of task, goal and role clarity respectively. We found a significant effect of task type on the perception of task clarity at the $p < .01$ level, across the 6 task type conditions; $F(5, 7002) = 6.176, p < .001$. Post-hoc comparisons using the Tukey HSD test indicated that the perception of task clarity in some task types was significantly poorer than others; as presented in Table 2.

Table 2: Post-hoc comparisons using the Tukey HSD test to investigate the effect of task types on *task clarity*. Comparisons resulting in significant outcomes are presented here. (* indicates $p < .05$ and ** indicates $p < .01$)

Task Type	M	SD	Comparison	Tukey HSD p-value
CA	3.75	.51	*CA vs SU*	0.011*
CC	3.76	.51	*CA vs VV*	0.004**
IA	3.74	.52	*CC vs SU*	0.046*
IF	3.77	.52	*CC vs VV*	0.020*
SU	3.82	.50	*IA vs SU*	0.001**
VV	3.82	.48	*IA vs VV*	0.001**

Table 3: Post-hoc comparisons using the Tukey HSD test to investigate the effect of task types on *goal clarity*. Comparisons resulting in significant outcomes are presented here. (* indicates $p < .05$ and ** indicates $p < .01$)

Task Type	M	SD	Comparison	Tukey HSD p-value
CA	3.76	0.52	*CA vs VV*	0.006**
CC	3.76	0.50	*CC vs VV*	0.004**
IA	3.74	0.51	*IA vs SU*	0.005**
IF	3.78	0.52	*IA vs VV*	0.001**
SU	3.82	0.51		
VV	3.83	0.50		

Table 4: Post-hoc comparisons using the Tukey HSD test to investigate the effect of task types on *role clarity*. Comparisons resulting in significant outcomes are presented here. (* indicates $p < .05$ and ** indicates $p < .01$)

Task Type	M	SD	Comparison	Tukey HSD p-value
CA	3.75	0.51	*CA vs SU*	0.030*
CC	3.76	0.50	*CA vs VV*	0.001**
IA	3.73	0.52	*CC vs SU*	0.048*
IF	3.78	0.51	*CC vs VV*	0.001**
SU	3.82	0.50	*IA vs SU*	0.001**
VV	3.84	0.48	*IA vs VV*	0.001**
			IF vs VV	0.043*

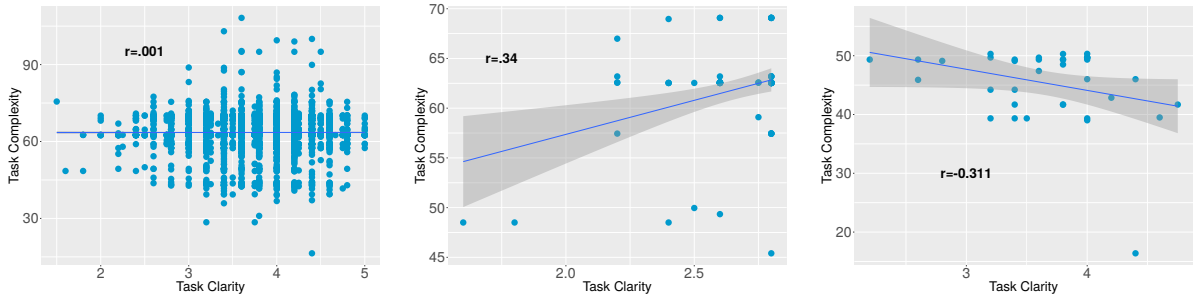
We also found a significant effect of task type on (i) the perception of goal clarity at the $p < .01$ level, across the 6 task type conditions; $F(5, 7002) = 5.918, p < .001$, and (ii) the perception of role clarity at the $p < .01$ level, across the 6 task type conditions; $F(5, 7002) = 8.074, p < .001$. Post-hoc comparisons using the Tukey HSD test (Tables 3 and 4) indicated that the perception of goal and role clarity in some task types was significantly poorer than others.

4.4 Task Clarity and Task Complexity

Recent work by Yang et al. modeled *task complexity* in crowdsourcing microtasks [37]. By using the task complexity predictor proposed by the authors, we explored the relationship between task clarity and task complexity. We found no significant correlation between the two variables across the different types of 7,100 tasks in our dataset (see Figure 5a). This absence of a linear relationship between task complexity and task clarity suggests that tasks with high clarity can still be highly complex or tasks with low clarity can have low task complexity at the same time.

We analyzed the relationship between *task clarity* and *complexity* across different types of tasks, and found that there is no observable correlation between the two variables across the different types of tasks. As can be observed from Figure 3, a majority of tasks are perceived to lie within the range of moderate to high clarity. We therefore further investigated tasks with low clarity or complexity.

4.4.1 Relationship in Tasks with Low Clarity. As shown earlier, task clarity was coherently perceived by workers. We reason that tasks corresponding to a clarity rating < 3 have relatively *low* clarity. We investigated the effect of task types on the relationship between task clarity and complexity in tasks with low clarity. Using Pearson's **r**, we found a weak positive linear relationship between the two variables in information finding (IF) tasks with low clarity (see Figure 5b); *N=80*, **r**=*.34*. This can be explained as a



(a) Overall relationship between *task clarity* and *task complexity*.

(b) Relationship in information finding (IF) tasks with low *task clarity*.

(c) Relationship in content access (CA) tasks with low *task complexity*.

Figure 5: Relationship between *task clarity* and *complexity*.

consequence of complex workflows required to complete some IF tasks, where high task complexity is concomitant with relatively high task clarity. Accordingly, in IF tasks with low clarity, task complexity accounted for 11.56% of the variance in task clarity (the coefficient of determination, R^2=*.1156, p<.01*). We did not find a significant relationship between the two variables in the low clarity subsets of other task types.

4.4.2 Relationship in Tasks with Low Complexity. Similarly, we consider tasks having a complexity score < 50 have relatively *low* complexity. We investigated the effect of task types on the relationship between task clarity and complexity in tasks with low complexity. Using Pearson's **r**, we found a weak negative linear relationship between the two variables in content access (CA) tasks with low complexity (see Figure 5c); *N=41*, **r**=*.311*. Thus, in CA tasks with low complexity, task clarity accounted for 9.67% of the variance in task complexity (the coefficient of determination, R^2=*.0967, p<.05*).

4.4.3 Discussion. The lack of linear correlation between clarity and complexity yields interesting observations. While surprising (intuitively, one might assume that a better task formulation – high clarity – would yield a lower complexity), this result is aligned with the classical theory on cognitive load, by Sweller and Chandler [35]. The theory postulates the presence of two sources of cognitive load: *intrinsic* and *extraneous*. Intrinsic cognitive load refers to the inherent difficulty in the content of presented material, which approximates task complexity in our context; extraneous cognitive load, on the other hand, refers to the organization and presentation of material, i.e. task clarity in our context. Sweller and Chandler suggest in their theory that, while the intrinsic cognitive load is unalterable, the extraneous cognitive load can either be increased because of inappropriate instructional design, or be reduced by well-structured presentation. We show that the theory can find application in microtask crowdsourcing, as tasks of similar complexity can either be of high clarity or low clarity.

When considering tasks of specific types, however, we find correlation could be established. Specifically, we find a negative correlation with content access (CA) tasks, thus suggesting that (poorly formulated) tasks asking workers to interact with on line content (e.g. watch a video, click a link) can be perceived more complex to execute. With information finding (IF) tasks, high task complexity maps with high clarity, thus suggesting that requests for complex finding and retrieval operations can be associated with clearer instructions. These results provide further insights into the relationship between task clarity and complexity, and call for further investigation.

5 PREDICTION OF TASK CLARITY

In this section we tackle **RQ3** and propose to model task clarity based on objective features that are extractable from tasks. We envision a system that could automatically predict task clarity and thus provides feedback to requesters on task design and to workers on task selection and execution. To test the feasibility of this idea, our study starts by designing task features that are potentially predictive for task clarity; we then build a predictive model to automatically learn task clarity based on these features.

5.1 Features Sets

We explore four classes of task features, namely: *metadata features*, *task type features*, *content features*, and *readability features*. In the following we provide a brief introduction to each feature class, and refer the readers to the companion page for a full description of the feature set.

Metadata Features are the task attributes associated with the definition of tasks when they are created. Typical metadata features include the number of `initial HITs`, attributes of the descriptions about desired activities to be performed by workers (e.g., `title length` and `description length`), the required qualification of workers (e.g., worker `location` and minimum `approval rate`), the estimated execution time (i.e. `allotted time`) and `reward`. These features characterize a task from different aspects that might be correlated with task clarity. For example, we assume that a longer description could entail more efforts from the requester in explaining the task.

Task Type Features categorize a task into one of the six task types defined by [14]. They are therefore high level features that comprehensively describe what knowledge is in demand. Through previous analysis, we have observed that task type has a significant effect on the perception of task clarity. We therefore assume that task type could be indicative of task clarity in prediction.

Content Features capture the semantics of a task. These features use the high-dimensional bag of words (BOW) representation. To maximize the informativeness of the content features while minimizing the amount of noise, one-hot (i.e. binary) coding was applied to the BOW feature of task title and keywords, while TF-IDF weighting was applied to the BOW feature of task description. It has been shown by research in related domains (e.g., community Q&A systems [36]) that the use of words is indicative of the quality of task formulation, therefore we are interested in understanding the effect of language use on workers' perception of task clarity.

Readability Features are by nature correlated with task clarity: tasks with higher readability are better formulated, and are thereby expected to have a higher clarity. We experiment with several widely used readability metrics in our clarity prediction task to understand their predictive power of task clarity. These include the use of long words (`long_words`), long sentence (`words_per_sentence`), the use of `preposition`, `nominalization`, and more comprehensive readability metrics such as `ARI`, `LIX`, and in particular, `Coleman_Liau`, which approximates the U.S. grade level necessary to comprehend a piece of text.

5.2 Prediction Results

Due to the high dimension of the content features (size of vocabulary = 10,879), we apply the Lasso method, which does feature selection and regression simultaneously. We adopt 5-fold cross-validation and mean absolute error (MAE) for evaluation. Table 5 shows the prediction results. The prediction on task clarity achieves a MAE of 0.4032 (SD = 0.0031). The relatively small error compared to the scale of ground truth (i.e. 1-5) indicates that task clarity can be predicted accurately. In addition, the small standard deviation shows that the prediction is robust across different tasks. Similar results also hold for the prediction of goal clarity and role clarity, which confirms our previous observation that both are highly correlated with the overall task clarity.

Table 5: Prediction results for Task Clarity, Goal Clarity and Role Clarity, shown by $\mu \pm \sigma$.

Clarity	Task Clarity	Goal Clarity	Role Clarity
MAE	0.4032±0.0031	0.4076±0.0067	0.4008± 0.0070

Predictive Features. In the following we analyze the predictive features selected by Lasso. Table 6 shows the features with positive and negative coefficients in the Lasso model after training for task clarity prediction, i.e. features that are positively and negatively correlated with task clarity. Similar observations can be obtained for predicting goal and role clarity.

With regard to metadata features, it can be observed that longer descriptions and more keywords are positively correlated with task clarity. This suggests that more description and keywords could potentially improve the clarity of task formulation. We also observe that the increased use of images, or less use of external links could enhance task clarity. These are reasonable, since intuitively, images can help illustrate task requirements, while external links would bring in extra ambiguity to task specification in the absence of detailed explanations.

Table 6: Predictive features for task clarity prediction.

Feat. Class	**Feat. w. Positive Coef.**		**Feat. w. Negative Coef.**	
	Feature	Coef.*	Feature	Coef.*
Metadata	`number_keywords`	0.719	`external_links`	-0.598
	`description_length`	0.295		
	`number_images`	0.071		
	`total_approved`	0.011		
Task Type	`VV`	0.434	`IA`	-0.922
	`SU`	0.413		
Content	keyword: `audio`	2.673	keyword: `id`	-2.658
	keyword: `transcription`	1.548		
	keyword: `survey`	1.178		
Readability	`preposition`	1.748	`ARI`	-1.982
	`GunningFogIndex`	1.467	`long_words`	-0.671
	`Coleman_Liau`	0.855	`syllables`	-0.478
	`words_per_sentence`	0.620	`nominalization`	-0.136
	`characters`	0.237	`pronoun`	-0.104
	`LIX`	0.150	`FleschReadingEase`	-0.075
			`RIX`	-0.038
	(all about title)		(all about title)	

* For the sake of comparison, each value is shown with original coefficient $\times 10^2$.

With regard to task type features, we find that tasks of type SU and VV are in general of higher clarity, while tasks of type IA are of lower clarity. This result confirms our previous findings.

With regard to content features, we observe that keyword features are more predictive than other types of content features (e.g. words in title or description). Predictive keywords include `audio`, `transcription`, `survey`, etc., which can actually characterize the majority of tasks in AMT. We therefore reason that workers' familiarity with similar tasks could enhance their perception of task clarity.

Finally, several interesting findings with regard to task readability are observed as follows. First, many types of readability scores are indicative of task clarity, indicating a strong correlation between task readability and task clarity. Second, compared with description or keyword readability, title readability is most predictive of task clarity. As an implication for requesters, putting efforts in designing better titles can improve task clarity. Third, we observe a positive correlation between task clarity and `Coleman_Liau`, which approximates the U.S. grade level necessary to comprehend the text. The increase of Coleman_Liau (i.e. more requirements on workersfi capability to comprehend the title) therefore does not lead to lower task clarity perceived by workers. The result is not surprising, given the demographic statistics of crowdworkers [13]. However, it raises questions on the suitability of this class of microtask crowdsourcing tasks for other types of working population.

On decomposing `Coleman_Liau` and exploring the effect of length of words (in terms of #letters) and length of sentences (in terms of #words), it can be observed that longer words (i.e., `long_words`) would decrease task clarity, while longer sentence (i.e., `words_per_sentence`) can enhance task clarity. This suggests that workers can generally comprehend long sentences, while the use of long words would decrease task clarity. This is consistent with our findings from **RQ1**, where workers identified difficult words as a factor that decreased task clarity and also suggested that tasks with difficult words are commonplace in the microtask crowdsourcing market. We also found a positive correlation between

preposition and task clarity, in contrast to the negative correlation between syllables (or nominalization) and task clarity. These results suggest that partitioning sentences with prepositions could increase task clarity, while complicating individual words decreases task clarity.

6 EVOLUTION OF TASK CLARITY

6.1 Role of Task Types

To address **RQ4**, we investigated the evolution of task clarity over time (see Figure 6). We found that there was no monotonous trend in the overall average task clarity over time, as shown in Figure 6a. We also investigated the effect of task type on the evolution of task clarity. We found no discernible trend in the evolution of task clarity of different types of tasks over the 12 month period considered in the dataset (Figure 6b). We conducted a one-way ANOVA to compare the effect of task type on the evolution of task clarity over time. We did not find a significant effect of task type on the evolution of task clarity at the $p < .05$ level, across the 6 task type conditions; $F(5, 66) = 0.081, p = .994$.

These findings suggest that the overall task clarity in the marketplace varies over time but does not follow a clear pattern. This can be attributed to the organic influx of new task requesters every month [13]. To identify whether the experience of task requesters plays a role in the evolution of task clarity, i.e., whether individual requesters deploy tasks with increasing task clarity over time we investigated the role of requesters in the evolution of task clarity.

6.2 Role of Requesters

Recent analysis of the AMT marketplace, revealed that there is an organic growth in the number of active requesters and a constant growth in the number of new requesters (at the rate of 1,000 new requesters per month) on the platform [13]. Poor task design leading to a lack of task clarity can be attributed to inexperienced requesters. To assess the role of requesters in the evolution of task clarity, we analyzed the evolution of task clarity of different types of tasks with respect to individual requesters.

We analyzed the distribution of unique requesters corresponding to the 7.1K tasks in our dataset. We found that a few requesters deployed a large portion of tasks, as depicted by the power law relationship in Figure 6c. We also found that over 40% of the requesters exhibited an overall average task clarity of $\geq 4/5$, and in case of nearly 75% of the requesters it was found to be over 3.5/5 (as presented in Figure 6d). We considered requesters who deployed ≥ 15 tasks within the 12-month period as being experienced requesters, and analyzed the relationship between the number of tasks they deployed with the corresponding overall task clarity. Using Pearson's **r**, we found a weak positive correlation between the average task clarity and the number of tasks deployed by experienced requesters (see Figure 6e); **r**= .28. Thus, the experience of requesters (i.e., the number of tasks deployed) explains over 8% of the variance in the average task clarity of tasks deployed by the corresponding requesters; the coefficient of determination, $R^2 = 0.081$.

Considering the requesters who deployed tasks during more than 6 months in the 12-month period, we investigated the overall change in terms of average task clarity of the tasks deployed from one month to the next. We measure the overall change in task clarity for each requester using the following equation.

$$\Delta TaskClarity_r = \frac{1}{n}\sum_{i=1}^{n}(TC_{i+1} - TC_i)$$

where, TC_i represents the average task clarity of tasks deployed by a requester in the month i, n is the total number of months during which requester r deployed tasks.

Figure 6f presents our findings with respect to the overall change in task clarity corresponding to such requesters. The size of the points representing each requester depict the number of tasks deployed by that requester. We did not find a significant effect of the number of tasks deployed by requesters on the magnitude of change in task clarity.

Based on our findings, we understand that the overall task clarity in the marketplace fluctuates over time. We found a weak positive linear relationship between the number of tasks deployed by individual task requesters and the associated task clarity over time. However, we did not find evidence that the magnitude of change in task clarity is always positive in case of experienced requesters.

6.3 Top Requesters

We note that the top-3 task requesters accounted for around 67% of the tasks that were deployed between Oct'13 to Sep'14. The requesters were found to be *SpeechInk*–1,061 tasks, *AdTagger*–944 tasks, and *CastingWords*–824 tasks. The evolution of task clarity of the tasks corresponding to these requesters over time is presented in the Figure 7 below.

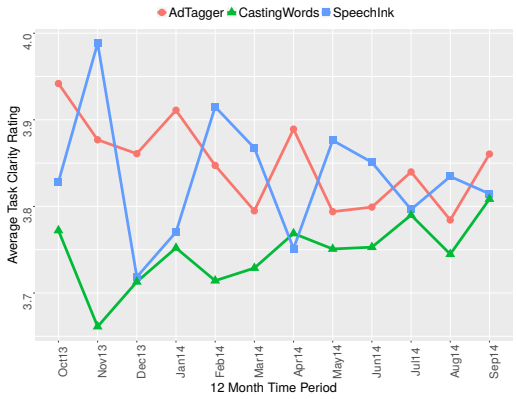


Figure 7: Top-3 task requesters w.r.t. the number of tasks deployed, and the evolution of their task clarity over time.

To understand the effect of the task requesters on the evolution of task clarity over time, we conducted a one-way between requesters ANOVA. We found a significant effect of task requesters on the evolution of task clarity across the three different requester conditions (SpeechInk, AdTagger, CastingWords) over the 12-month period; $F(2, 33) = 11.837, p < .001$. Post-hoc comparisons using the Tukey HSD test revealed that the evolution of task clarity corresponding to tasks from *SpeechInk* and *AdTagger* were significantly different in comparison to *CastingWords*.

We observe a gradual increase in the task clarity of *CastingWords* tasks over time in contrast to the other two top requesters.

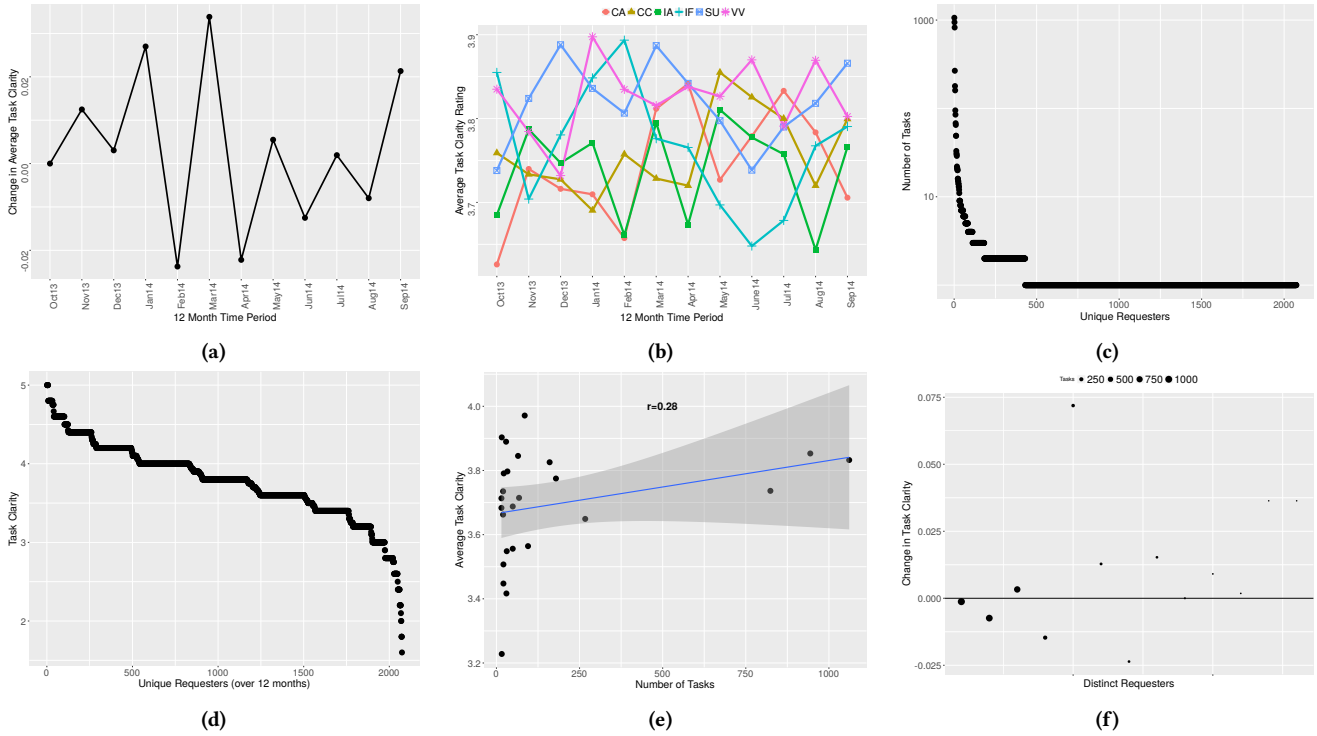


Figure 6: (a) Evolution of overall task clarity and (b) with respect to different types of tasks from Oct'13-Sep'14, (c) distribution of tasks corresponding to requesters who deployed them, (d) distribution of the average task clarity of tasks corresponding to distinct requesters across the 12 months, (e) relationship between the average task clarity and the number of tasks deployed by experienced requesters, (f) $\Delta TaskClarity$ of requesters who deployed tasks during more than 6/12 months in our dataset.

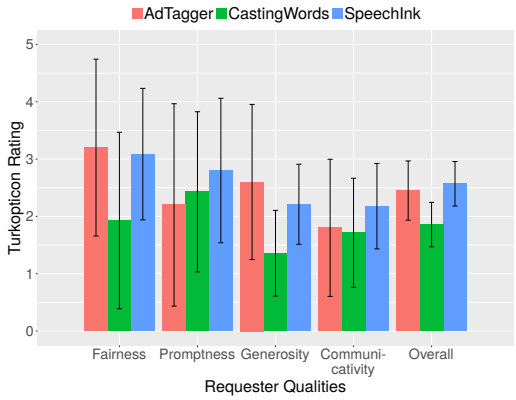


Figure 8: Average Turkopticon ratings of the top requesters from Oct'13–Sep'14.

In the context of these requesters and the time period of Oct'13-Sep'14, we explored the Turkopticon ratings [21] corresponding to the requesters. Turkopticon collects ratings from workers on the following qualities : *fairness* of a requester in approving/rejecting work, *communicativity*– the responsiveness of a requester when contacted, *generosity*– quality of pay with respect to the amount of time required for task completion, *promptness* of the requester in approving work and paying the workers. Figure 8 presents a comparison of the Turkopticon ratings of the 3 requesters for each of the four qualities. We note that *SpeechInk* received consistently better ratings across all qualities within the given period. This coincides with the relatively higher task clarity of *SpeechInk* (*M=3.83, SD=0.47*) tasks when compared to *CastingWords* (*M=3.73, SD=0.48*) tasks over the 12 months (see Figure 7). A two-tailed T-test revealed a significant difference in the task clarity between *SpeechInk* and *CastingWords*; $t(1883)=18.43$, $p < .001$. We did not find ratings of tasks deployed by *AdTagger* on Turkopticon during the time period considered. However, we present a comparison based on the ratings received by *AdTagger* prior to Oct'13. Once again, in comparison to *CastingWords* we note that the higher overall quality ratings of *AdTagger* on Turkopticon coincide with the higher task clarity over the 12 months (*M=3.85, SD=0.48*); $t(1766)=25.23$, $p < .001$.

Through our findings it is clear that task clarity is not a global, but a local property of the AMT marketplace. It is influenced by the actors in the marketplace (i.e., tasks, requesters and workers) and fluctuates with the changing market dynamics.

7 CONCLUSIONS AND FUTURE WORK

In this paper we examined *task clarity*, an important, yet largely neglected aspect of crowdsourced microtasks. By surveying 100 workers, we found that workers confront unclear tasks on a regular basis. They deal with such tasks by either exerting extra effort to overcome the suboptimal clarity, or by executing them without a clear understanding. Poor task formulation thereby greatly hinders

the progress of workers' in obtaining rewards, and in building up a good reputation.

To better understand how clarity is perceived by workers, we collected workers' assessments for 7.1K tasks sampled from a 5 years worth dataset of the AMT marketplace. With an extensive study we revealed that clarity is coherently perceived by workers, and that it varies by the task type. In addition, we found compelling evidence about the lack of direct correlation between clarity and complexity, showing the presence of a complex relationship that requires further investigation. We proposed a supervised machine learning model to predict task clarity and showed that clarity can be accurately predicted. We found that workers' perception of task clarity is most influenced by the number of keywords and title readability. Finally, through temporal analysis, we show that clarity is not a macro-property of the AMT ecosystem, but rather a local property influenced by tasks and requesters.

In conclusion, we demonstrated the importance of clarity as an explicit property of microwork crowdsourcing tasks, we proposed an automatic way to measure it, and we unveiled interesting relationships (or lack thereof) with syntactical and cognitive properties of tasks. Our findings enrich the current understanding of crowd work and bear important implications on structuring workflow. Predicting task clarity can assist workers in task selection and guide requesters in task design. In the imminent future, we will investigate the impact of task clarity in shaping market dynamics such as worker retention versus dropout rates.

Acknowledgments

This research has been supported in part by the European Commission within the H2020-ICT-2015 Programme (Analytics for Everyday Learning – AFEL project, Grant Agreement No. 687916), the Dutch national e-infrastructure with the support of SURF Cooperative, and the Social Urban Data Lab (SUDL) of the Amsterdam Institute for Advanced Metropolitan Solutions (AMS).

REFERENCES

[1] Omar Alonso and Ricardo Baeza-Yates. 2011. Design and implementation of relevance assessments using crowdsourcing. In *ECIR*. Springer, 153–164.
[2] Omar Alonso, Catherine Marshall, and Marc Najork. 2014. *Crowdsourcing a subjective labeling task: a human-centered framework to ensure reliable results.* Technical Report. MSR-TR-2014-91.
[3] Janine Berg. 2016. Income security in the on-demand economy: findings and policy lessons from a survey of crowdworkers. *Comparative Labor Law & Policy Journal* 37, 3 (2016).
[4] Hein Broekkamp, Bernadette HAM van Hout-Wolters, Gert Rijlaarsdam, and Huub van den Bergh. 2002. Importance in instructional text: teachers' and students' perceptions of task demands. *Journal of Educational Psychology* 94, 2 (2002), 260.
[5] Francisco Cano and María Cardelle-Elawar. 2004. An integrated analysis of secondary school studentsfi conceptions and beliefs about learning. *European Journal of Psychology of Education* 19, 2 (2004), 167–187.
[6] Jeanne Sternlicht Chall and Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula.* Brookline Books.
[7] Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of current and future research. *ITL-International Journal of Applied Linguistics* 165, 2 (2014), 97–135.
[8] Kevyn Collins-Thompson and James P Callan. 2004. A language modeling approach to predicting reading difficulty.. In *HLT-NAACL*. 193–200.
[9] Scott A Crossley, Kristopher Kyle, and Danielle S McNamara. 2015. The tool for the automatic analysis of text cohesion (TAACO): automatic assessment of local, global, and text cohesion. *Behavior research methods* (2015), 1–11.
[10] Tove I Dahl, Margrethe Bals, and Anne Lene Turi. 2005. Are students' beliefs about knowledge and learning associated with their reported use of learning strategies? *British journal of educational psychology* 75, 2 (2005), 257–273.
[11] Edgar Dale and Jeanne S Chall. 1949. The concept of readability. *Elementary English* 26, 1 (1949), 19–26.
[12] Orphée De Clercq, Véronique Hoste, Bart Desmet, Philip Van Oosten, Martine De Cock, and Lieve Macken. 2014. Using the crowd for readability prediction. *Natural Language Engineering* 20, 03 (2014).
[13] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G Ipeirotis, and Philippe Cudré-Mauroux. 2015. The Dynamics of Micro-Task Crowdsourcing: The Case of Amazon MTurk. In *WWW*. International World Wide Web Conferences Steering Committee, 238–247.
[14] Ujwal Gadiraju, Ricardo Kawase, and Stefan Dietze. 2014. A taxonomy of microtasks on the web. In *Hypertext*. ACM, 218–223.
[15] Ujwal Gadiraju, Ricardo Kawase, Stefan Dietze, and Gianluca Demartini. 2015. Understanding malicious behavior in crowdsourcing platforms: the case of online surveys. In *CHI*. ACM, 1631–1640.
[16] Catherine Grady and Matthew Lease. 2010. Crowdsourcing document relevance assessment with mechanical turk. In *HLT-NAACL workshop on creating speech and language data with Amazon's mechanical turk*. Association for ComputationalLinguistics, 172–179.
[17] Arthur C Graesser, Danielle S McNamara, Max M Louwerse, and Zhiqiang Cai. 2004. Coh-Metrix: analysis of text on cohesion and language. *Behavior research methods, instruments, & computers* 36, 2 (2004), 193–202.
[18] Allison Hadwin. 2006. Student task understanding. In *Learning and Teaching Conference. University of Victoria, Victoria, British Columbia, Canada.*
[19] AF Hadwin, M Oshige, M Miller, and P Wild. 2009. Examining student and instructor task perceptions in a complex engineering design task. In *international conference on innovation and practices in engineering design and engineering education. McMaster University, Hamilton, ON, Canada.*
[20] T Hoßfeld, Raimund Schatz, and Sebastian Egger. 2011. SOS: The MOS is not enough!. In *QoMEX*. IEEE, 131–136.
[21] Lilly C Irani and M Silberman. 2013. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *CHI*. ACM, 611–620.
[22] Diane Lee Jamieson-Noel. 2004. *Exploring task definition as a facet of self-regulated learning*. Ph.D. Dissertation. Faculty of Education-Simon Fraser University.
[23] Rohit J Kate, Xiaoqiang Luo, Siddharth Patwardhan, Martin Franz, Radu Florian, Raymond J Mooney, Salim Roukos, and Chris Welty. 2010. Learning to predict readability using diverse linguistic features. In *ACL*. Association for Computational Linguistics, 546–554.
[24] Shashank Khanna, Aishwarya Ratan, James Davis, and William Thies. 2010. Evaluating and improving the usability of Mechanical Turk for low-income workers in India. In *DEV*. ACM, 12.
[25] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel.* Technical Report. DTIC Document.
[26] Aniket Kittur, Ed H Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *CHI*. ACM, 453–456.
[27] Aniket Kittur, Jeffrey V Nickerson, Michael Bernstein, Elizabeth Gerber, Aaron Shaw, John Zimmerman, Matt Lease, and John Horton. 2013. The future of crowd work. In *CSCW*. ACM, 1301–1318.
[28] Lieve Luyten, Joost Lowyck, and Francis Tuerlinckx. 2001. Task perception as a mediating variable: A contribution to the validation of instructional knowledge. *British Journal of Educational Psychology* 71, 2 (2001), 203–223.
[29] David Malvern and Brian Richards. 2012. Measures of lexical richness. *The Encyclopedia of Applied Linguistics* (2012).
[30] Catherine C Marshall and Frank M Shipman. 2013. Experiences surveying the crowd: Reflections on methods, participation, and reliability. In *WebSci*. ACM, 234–243.
[31] Emily Pitler and Ani Nenkova. 2008. Revisiting readability: A unified framework for predicting text quality. In *EMNLP*. Association for Computational Linguistics, 186–195.
[32] Presentacion Rivera-Reyes. 2015. Students' task interpretation and conceptual understanding in electronics laboratory work. (2015).
[33] Libby O Ruch and Rae R Newton. 1977. Sex characteristics, task clarity, and authority. *Sex Roles* 3, 5 (1977), 479–494.
[34] Aaron D Shaw, John J Horton, and Daniel L Chen. 2011. Designing incentives for inexpert human raters. In *CSCW*. ACM, 275–284.
[35] John Sweller and Paul Chandler. 1994. Why some material is difficult to learn. *Cognition and instruction* 12, 3 (1994), 185–233.
[36] Jie Yang, Claudia Hauff, Alessandro Bozzon, and Geert-Jan Houben. 2014. Asking the right question in collaborative q&a systems. In *Hypertext*. ACM, 179–189.
[37] Jie Yang, Judith Redi, Gianluca Demartini, and Alessandro Bozzon. 2016. Modeling task complexity in crowdsourcing. In *HCOMP*. AAAI, 249–258.