SCAFFOLDING CROWD WORK

—AN EMPIRICAL ANALYSIS OF ITS EFFECTS ON QUALITY OF WORK

Monika Streuer

Supervisor: Prof. Magie Hall
Examiner: Prof. Alf Westelius
Contents

List of figures ........................................................................................................................................iv
List of tables ........................................................................................................................................v

1. Introduction ................................................................................................................................... 9
   1.1. Motivation .................................................................................................................................. 9
   1.2. Objectives .................................................................................................................................. 11
   1.3. Research Questions .................................................................................................................. 11
   1.4. Structure of the Thesis ............................................................................................................. 12

2. Background ..................................................................................................................................... 13
   2.1. Crowd Work ............................................................................................................................. 13
       2.1.1. Related Academic Concepts ............................................................................................. 13
       2.1.2. The Concept of Crowd Work ............................................................................................ 15
       2.1.3. Benefits and Challenges of Paid Crowd Work ................................................................. 17
       2.1.4. Online Labor Platforms .................................................................................................... 18
   2.2. Learning .................................................................................................................................... 22
       2.2.1. Formal Education ............................................................................................................... 22
       2.2.2. Learning from Feedback in Crowdsourcing ..................................................................... 24
   2.3. Improving Work Quality ......................................................................................................... 25
   2.4. Applying Scalable Methods to Crowds .................................................................................... 26
       2.4.1. Examples ............................................................................................................................ 26
       2.4.2. Motivating Task Rationale ............................................................................................... 27
       2.4.3. Rubrics ................................................................................................................................ 27
       2.4.4. Step-by-step Instructions ................................................................................................... 28

3. Own Approach ................................................................................................................................. 29
   3.1. Participants ............................................................................................................................... 30
   3.2. Subject of Study – the Tasks .................................................................................................... 31
       3.2.1. Scaffolding .......................................................................................................................... 32
       3.2.2. Writing task ....................................................................................................................... 35
       3.2.3. Slide Design Task ............................................................................................................. 38
   3.3. Procedure .................................................................................................................................. 40
       3.3.1. Data Collection .................................................................................................................. 40
       3.3.2. Evaluation of Submissions ................................................................................................. 40
       3.3.3. Payment ............................................................................................................................. 42
   3.4. Measures ..................................................................................................................................... 42
       3.4.1. Independent Variables ....................................................................................................... 43


3.4.2. Dependent Variables

4. Evaluation

4.1. Data Analysis

4.2. Data Interpretation

4.2.1. The Performance of Micro-Task Workers is Being Underestimated

4.2.2. Scaffolding, Leads to Better Performance

5. Discussion

5.1. Results of the Study

5.2. Discussion of the Findings

5.3. Implications

5.4. Critical Evaluation of the Study

6. Conclusions and Future Outlook

7. Declaration about the thesis

Appendix

Appendix A

Appendix B

Reference List
List of figures

Figure 1: Structure of the thesis ................................................................. 12
Figure 2: Academic concepts related to crowd work .................................. 15
Figure 3: The basic functionality of online labor markets .......................... 19
Figure 4: Mapping complexity vs. payment in a crowd work design space ... 21
Figure 5: Real work examples from MTurk .................................................. 22
Figure 6: Real work examples from Upwork .............................................. 22
Figure 7: Study design ............................................................................. 30
Figure 8: Example of a well-designed slide .............................................. 33
Figure 9: The writing task under original condition .................................. 35
Figure 10: Extract of the scaffolded writing task ...................................... 37
Figure 11: Slide-design task under original condition ............................... 38
Figure 12: Extract of the scaffolded slide-design task ............................... 39
Figure 13: List of criteria for evaluating the writing task ............................ 41
Figure 14: List of criteria for evaluating the slide-design task .................... 42
Figure 15: Examples of submissions - slide design task ........................... 48
List of tables

Table 1: Main and interaction effects between the three independent variables...........44
Table 2: Summary of work performance in all six possible settings................................45
Table 3: Summary of the multi-linear regression model.............................................45
Abstract

Workers’ level of skill on online labor platforms varies greatly. To manage quality, requesters often decompose complex jobs into simple, repetitive, micro-tasks that pay in average $0.10 per task. Within the current system, workers waste a significant amount of time on finding micro-tasks while requesters still criticize the performance, struggle with decomposing complex work into micro-tasks, allocating these tasks appropriately and reintegrating the outcome to a final result. This paper explores how we can apply scaffolding approaches — examples, rubrics, task rationale, and step-by-step instructions — to improve work quality and enable workers to perform more complex tasks. In a between-subjects study, novice workers from a micro-task market performed tasks (writing product reviews and designing a slide deck) selected from a professional contractor market. Participants received either the original or a scaffolded task description. Blind-to-condition experts judged the performance of submissions. We found that scaffolding the crowd led workers to perform significantly better than workers without scaffolding. Moreover, a follow-up analysis shows that scaffolding micro-task workers results in work that is on par with workers from the high-pay contractor platform. Concluding, it is being discussed if, how and by who scaffolding in a crowdsourcing environment can and should be implemented.
Acknowledgments

The name that graces the cover of this thesis is mine and mine alone. However, having said this, there are a number of people without whose support and guidance I would not have been able to write it. This acknowledgments section is my way of expressing appreciation for these very people – advisors, colleagues, friends and family – who inspired my work both in the past and will continue to do so in the future.

First and foremost, my academic advisors – Markus Krause, Steven Dow and Magie Hall – have been a constant source of support and lighting beacon throughout the entire process of researching and writing. My eternal gratitude for your time, your contributions, your ideas and constant valuable feedback. Your enthusiasm inspired me and I hope to not just be able to prolong our working relationship but also to make you proud as I embark on the next step in my academic career.

Special thanks to the DesignLab and ProtoLab at UC San Diego for hosting me as a visiting researcher. The synergistic and collaborative atmosphere of both labs was a fruitful and inspiring resources not just for the development of my research, but also for my personal development as a novice researcher. In particular, my students Xinran Wie and Se One Park, you were wonderful, and I highly appreciated your time and effort in contributing to my study.

On a more personal level, I am indebted to Jan Tesch, who is not just a colleague, but also a wonderful friend whose time, constructive feedback and constant encouragement allowed me to follow and realize my dreams (no matter how impossible they might have seemed at some point).

To my friends, Maja Stich – checking on me every single day of my stay in the US, Fabian Telschow – working night shifts with me, Pascal Kraft – making me smile, Rafael Lorenz – giving me constant constructive feedback, Liz Cahill – my creative muse, Cecilia Schellhaas - my oldest friend, Mieke Vandenbroucke – role model, Philip Stroisch – highly appreciated distraction and many more, I am afraid despite my effort, I may not recall each face that went with each moment of support, I do not have the words to tell you how much I value you being part of my life.
Last but not least, writing a thesis is not an easy and straightforward endeavor; I am blessed by the fact that the Schiöberg family not only provided a roof over my head, but also because of their kindhearted nature became a second family during my stay in Berkeley;

My parents and grandmother back home in Germany, thank you for not only making my stay in the US possible, but also for always encouraging me to push myself to grow into the strong woman that you raised.

Thank you!

Monika Streuer
1. Introduction

The concept of crowdsourcing has already existed for centuries (Longitude Price; Toyota Motor Corporation; Winchester 1999) but it first became popular in 2006 when the term was coined by Howe (2006). Since then, its range of application and popularity have been growing constantly. The large number of users, requesters as well as workers, indicates its importance for the labor markets of the 21st century (Kassi & Lehdonvirta 2016). Unfortunately, neither requesters nor workers are entirely satisfied with the infrastructure the online labor market developed and the consequences for work and its outcome that go along with it (Irani & Silberman 2013; Harris 2014; Vakharia & Lease 2015). This thesis focuses on the issues of low-end micro-task online labor platforms investigating the possibility of scaffolding as means to enhance performance and lower the need for decomposition of work tasks on micro-task platforms.

1.1. Motivation

Outsourcing work to a generally large, geographically distributed network of people in the form of an open call provides quick access to human intelligence at reasonable cost, making it very attractive to commercial vendors (Surowiecki 2005; Amazon Mechanical Turk; Upwork; freelancer; Innocentive; 99designs; Chaordix; OneSpace). Hence, a wide range of platforms such as Amazon’s Mechnical Turk (MTurk), Upwork, freelancer, Innocentive, 99designs, etc. arose, specializing in connecting businesses looking for a workforce outside their own corporation with online workers seeking remote work opportunities. The ecosystem of online labor platforms can be characterized by the tasks’ complexity, degree of task decomposition and thus payment per task as the differentiating factors spanning a wide spectrum from higher pay, more complex, continuous contract work (offered for example on Upwork) and the lower pay, more repetitive, occasional micro-task work (offered on platforms such as MTurk). This spectrum is based on the development of two distinct and fundamentally different approaches requesters follow.

There are employers that strongly depend on the field specific knowledge of specialists and hence hire professional online workers, who they transfer the responsibility for a particular project with all sub-tasks that it involves to when signing a work contract. This
work normally continues for a longer period of time allowing workers to focus on the project and guarantees at the same time constant income for the time of being engage in the project. The opposite approach is represented by employers, who chose to utilize the most possible manpower and at the same time avoid failure based on a single human being. This is being realized by targeting the lowest common denominator of the workforce. These requesters decompose tasks to small units that require only basic cognitive skills (Quinn & Bederson 2011). This approach fostered the development of micro-task platforms such as MTurk, which hosts task that mostly pay $0.10/task or less (Hitlin 2016) with the consequence that workers on micro-task platforms, who are per se poorly compensated (Boonthum-Denecke 2011; Ipeirotis 2010b; Ross et al. 2010; Agrawal et al. 2015), must invest a significant amount of their time in finding fitting tasks and applying for them instead of completing work and making money.

Currently over 500,000 workers worldwide are active on MTurk (Hitlin 2016). A large portion of this group is dealing with the highlighted issue on a regular basis. 63% of the MTurk workers use this platform daily to complete tasks, another 32% at least once a week (Hitlin 2016), making MTurk a source for additional or in some cases even main income.

While work on platforms situated at the lower end of the spectrum is causing dissatisfaction for workers as explained above, it at the same time troubles requesters, who are the initiators of this very phenomenon of micro-task labor platforms. Often requesters claim the work on micro-task platforms to be of poor quality (Wais et al. 2010; Calvo et al. 2015, p.385). Additionally, they are facing challenges that originate from decomposing tasks. To receive a valuable outcome the requesters must a) choose the appropriate granularity to divide a job into a micro-task b) integrate the results into a final product (Becker & Murphy 1992; Benkler 2002) and c) subsequently, allocate the task to workers in a manner that best matches the skills needed, which requires finding a suitable platform hosting workers qualified for the job (Durward, Blohm & Leimeister 2016). (Kittur et al. 2013)
1.2. Objectives

To put an end to this inefficiency and struggling of both, workers and requesters, and to develop a sustainable infrastructure that transforms online work on micro-task platforms into a viable marketplace for satisfactory results, this research explores possibilities how to enhance performance on micro-task labor platforms and lower the need for decomposition of tasks.

Existing approaches such as the formal concept of learning (off- and online) for acquiring more knowledge as well as feedback and rubrics for achieving higher quality of work require time investment and commitment by workers and/or requesters. Crowdsourcing needs a more scalable approach to improve the capabilities of micro-task workers.

Drawing on the theory of “zone of proximal development” (Vygotsky 1980) as well as scaffolding, which both encourage giving novices support to solve tasks, (van der Stuyf 2002; Hammond & Gibbons 2005; Reiser 2004; Hogan & Pressley 1997), I was given the task to examine scaffolding task descriptions as means for enhancing performance and for avoiding task decomposition. The aim of this thesis was to explore the influence of providing appropriate guidance via scaffolding task descriptions using the (ex ante given) methods: example, rubrics, motivating task rationale, and step-by-step instructions on performance and ability of solving more complex tasks (as provided on higher-end platforms).

1.3. Research Questions

Toward the stated goal of improving work quality and enabling workers to complete more complex jobs, this research explores the following research questions:

1. Can scaffolding based on learning science techniques enable novices to perform more complex tasks?
2. Can task descriptions enhanced by using scaffolding enable novices to achieve higher quality of work?
3. Which learning techniques do novices find most helpful?
1.4. Structure of the Thesis

This thesis is structured into 6 chapters illustrated in Figure 1. The introduction following is the background for chapter 2. First it explains the relevant terms and concepts used for the thesis and points out the weaknesses of existing approaches. Then it provides an introduction of related concepts and a general definition of crowd work, and elaborates on the way crowd work functions, how it is being practiced and what benefits and drawbacks go along with it. Then, existing concepts for handling the emerged issues are explained. Furthermore, the theoretical foundation for the approach proposed in this thesis is introduced. The study itself including the choice of participants and their recruitment, the subject of study, the exact procedure the study follows as well as the measures used are explained in detail in chapter 3. Chapter 4 contains the presentation and evaluation of the data gathered. The results of this study with its implications and limitations are critically discussed in chapter 5. Chapter 6 concludes the thesis and reveals an outlook for future research.

Figure 1: Structure of the thesis
2. Background

Before the research methodology of this study is explained in detail, the background chapter gives an overview and summarizes the existing academic efforts within this field pointing out weaknesses which eventually lead to the specific research questions that are being explored within this study.

First, related academic concepts and the general concept of crowd work will be introduced, providing more details on the benefits and challenges going along with crowd work and the current state of online labor markets. Then, an overview of the existing approaches and past efforts dealing with how to enhance performance and the ability of solving more complex tasks in an online environment is given. Weaknesses of current approaches are pointed out leading to the theoretical foundation for this research that tries to use current knowledge and extend it. Lastly, the research questions are being derived and presented.

2.1. Crowd Work

2.1.1. Related Academic Concepts

While this work focuses on paid crowd work, academia has given rise to similar concepts such as human computation, social computing, and collective intelligence that are closely related to it, to some extent are the foundation for crowd work. These concepts are briefly introduced in the following section to provide and understanding of where crowd work sits in the academic field.

Human Computation

The term human computation was established in 2005 by Van Ahn (2005), inspired by Naor’s work on identification via the Turing test (1996). It refers to the use of online human participation as processing power in the computational process, for problems computers cannot yet solve autonomously.
However, the first discussion of this can be found as early as 1935 when the Langley Memorial Aeronautical Laboratory - at that time the main research center for the National Advisory Committee for Aeronautics (NACA)- offered computing as a profession (McLennan & Gainer 2012). The computing tasks, varying between different working units, mainly involved: reading film, running calculations, and plotting data were carried out by humans\(^1\). The idea of computers and humans interacting and replacing each other’s work was first mentioned by Turing as early as 1950 (Turing 1950). In 1960 Licklider envisioned a co-working-symbiosis of man and computers (Licklider 1960). However, this field has only just recently received academic attention (Quinn & Bederson 2011; Horvitz & Paek 2007; Quinn & Bederson 2011; Law & Ahn 2011). No general definition of collective intelligence has been established (yet), however most authors agree on the following aspects describing it (Quinn & Bederson 2011):

- The problems fit the general paradigm of computation, and as such might someday be completely solvable by computers.
- The human participation is directed by the computational system or process. (This is discussed more below.)

### Social Computing

Social computing is a novel and emerging computing paradigm involving computer and social science. By computing shifting to the edges of the network and by individuals becoming empowered to use the web, a new field of research developed exploring humans in a social role where communication is mediated by technology. Using technologies such as blogs, wikis and online communities as for example facebook, youtube, twitter, etc. constantly generates data on social behaviors on different media and platforms which social computing analyzes and reproduces as intelligent and applicable

\(^1\) The working units carrying our manual calculations and solving mathematical equations were called computers. This was the inspiration for the naming of electronic computers.
knowledge. (Parameswaran & Whinston 2007; Tavakolifard & Almeroth 2012; Quinn & Bederson 2011)

**Collective Intelligence**

Collective intelligence embraces a wide field of applications and realizations. Computer science mostly relates to Malone’s (Malone, Laubacher & Dellarocas 2009) definition of “groups of individuals doing things collectively that seem intelligent”. Hence, as illustrated in Figure 2, it is a superset of human computation, social computation and crowd work.

![Figure 2: Academic concepts related to crowd work](image)

### 2.1.2. The Concept of Crowd Work

Crowd work, also known as crowd sourcing and crowd employment is a concept that was in use long before the current online medium was invented. An early example is Toyota, who in 1936 held a contest to find a new brand logo receiving over 27000 submissions (Toyota Motor Corporation).

However, the term crowdsourcing became first popular in 2006 when introduced by Howe. In his article “The Rise of Crowdsourcing” he described a wide range of tasks newly enabled by internet connectivity (Howe 2006). Ever since the term crowd work was used to describe a sourcing model in which specific work tasks are outsourced via advanced internet technologies to an virtual online crowd (Howe 2006; Brabham 2010).
Though, a universal definition for crowdsourcing does not exist, Estellés-Arolas and González-Ladrón-de-Guevara (Estellés-Arolas & González-Ladrón-de-Guevara 2012), conducted a study analyzing common and disparate elements of definitions of crowdsourcing in academic publications identifying the following characteristics for a collective, consensus-based definition of the term:

“(a) there is a clearly defined crowd;
(b) there exists a task with a clear goal;
(c) the recompense received by the crowd is clear;
(d) the crowdsourcer is clearly identified;
(e) the compensation to be received by the crowdsourcer is clearly defined;
(f) it is an online assigned process of participative type;
(g) it uses an open call of variable extent;
(h) it uses the Internet.” (Estellés-Arolas & González-Ladrón-de-Guevara 2012, p. 197)

Crowd work is closely related to outsourcing and open sourcing. This mode of sourcing externalizes work to a work force outside the business (Kittur et al. 2013). However, while classic outsourcing assigns specific tasks to pre-defined actors (Quinn & Hilmer 1994), crowd sourcing relies on a widely unknown and distributed mass of people that solve the task either individually or by cumulative effort (Starbird 2012). The key difference to open sourcing is the skill sets crowd work utilizes (Boudreau & Lakhani 2013). Open source projects mainly require programming skills. Crowd work extends this concept by taking advantage of any skills that contribute to successfully doing work online, especially human intelligence that cannot be fully replaced by artificial solutions (Starbird 2012). Crowdsourced tasks can be as large as co-developing a software or creating an online encyclopedia and as small as counting pictures on a web-site or marking the difference between two pages.

Generally, crowd work can be categorized into voluntary community effort and paid work, but in the past, it has been used in a semi-voluntary manner. Voluntary crowd work is an individual online contribution intended to benefit other individuals or groups (Baruch, May & Yu 2016). It is often used for wikis (Bryant, Forte & Bruckman 2005),
games with a purpose (Ahn & Dabbish 2004), captchas (Ahn et al. 2003), and citizen science (Cooper et al. 2010; Savage 2012). With a rising number of voluntary participants scale effects can be achieved leading to a well-developed and highly valuable outcome for the community. Prominent examples are:

- Wikipedia – an online encyclopedia created and edited by volunteers and hence free of charge for users
- News during Crisis – the crowd providing information (words and images) to helpers and family members
- Open Source Projects such as Linux (computer operating system) and Firefox (browser) providing free software to users

Some businesses developed business models based on semi-voluntary crowd work. Duo Lingo (Duolingo), for example, is an online language learning app where novices practice their skills by solving increasingly complex vocabulary and grammar tasks. Initially the app was developed solely for learning purposes, but Duo Lingo extend their business by adding translation services. As part of the novices’ practice they were translating real tasks for free, which Duo Lingo then sold for money. LevelUp for Photoshop followed the same concept letting novices do real design tasks during their online learning process. (Dontcheva et al. 2014; Simonite 2013)

A significant share of online work, which this research focuses on, is paid crowd work. Requesters, who can be organization as well as individuals, make work tasks available online on online labor platforms where crowd workers complete them for compensation. A more detailed explanation is provided further on, when online labor platforms are being introduced.

2.1.3. Benefits and Challenges of Paid Crowd Work

Paid crowd work provides many positive aspects and advantages for both, workers and requesters. Crowd workers can work remotely without the need of commuting to work or transferring to a different place, they are (mostly) free to choose the time when they work as well as the amount of work they would like to take on (Kittur et al. 2013). Furthermore,
workers can apply for tasks involving content they would enjoy working on. For requesters crowd work gives quick access to a globally distributed network of people increasing the speed and flexibility businesses have while allowing diversity, all at reasonable cost (Prpi, Taeihagh & Melton 2015; Buettner 2015; Durward, Blohm & Leimeister 2016).

However, besides benefits crowd work holds challenges as well. The flexibility workers have concerning the amount and nature of tasks they can complete can become a burden if workers cannot find enough work, or if they must do poorly compensated work that additionally may not be within their area of interest. Furthermore, crowd work is a form of employment that is based on individual projects rather than on continuous contracts. Workers for whom crowd work is a primary source of income must deal with a high degree of uncertainty (Felstiner 2011). Due to the lack of a regular working contract they cannot claim any additional benefits (health insurance, retirement payment, etc.) offered to regular workers (Kittur et al. 2013).

Requesters as well face several concerns. First, receiving the right expertise may be difficult because there are fewer skilled workers than unskilled ones. Second, past crowdsourcing research as well as practice have strongly focused on increasing the degree of task-decomposition to target the lowest common dominator in the workforce (Kittur et al. 2013). This fostered a culture that goes along with additional effort and potential for failure. Requesters must choose the appropriate granularity to divide a job into a micro-task and then integrate the results into a final product (Becker & Murphy 1992; Benkler 2002). Subsequently, tasks must be allocated to workers in a manner that best matches the skills needed, which requires finding a suitable platform hosting workers qualified for the job. (Durward, Blohm & Leimeister 2016)

2.1.4. Online Labor Platforms

Online labor platforms are hosts for organizations seeking a workforce outside their own corporation and individuals looking for remote casual work (Horton 2010). They provide an organizational frame for connecting supply and demand in terms of online work.
The procedure of online work may vary slightly depending on the hosting platform, however, the basic concept remains the same as explained in the following and summarized in Figure 3.

![Figure 3: The basic functionality of online labor markets](image)

Organizations as well as individuals must sign in to a platform and can then immediately post work tasks as requesters. Each online labor platform has an individual set of questions that helps the requester to categorize the task correctly. Most platforms additionally require requesters to define the topic of the task, its difficulty level, the time required to solve it, or a deadline when the solution is due, and the compensation for the solution. Once the framework for the task is set, the requester can insert the task and post it. The task then appears in the pre-defined category at the platform and can be viewed by workers. Workers, who would like to do the task can apply for it. Mostly they are required to provide a curriculum vitae, credentials and work samples. The requester decides which applicant will be solving the tasks by accepting the application. The procedure of solving a task is often iterative. Preliminary work is reviewed by requesters, feedback and additional input is given until the requester is fully satisfied with the solution. The work is finished once the requester accepts the submitted work and initiates the payment for the task. (Hitlin 2016)
While the basic concept of how online labor platforms function is very similar the spectrum of their work is highly differentiated. The ecosystem of online labor platforms is characterized by the task’s complexity, degree of task decomposition and thus payment per task, still partly interdependent factors, spanning a spectrum from higher pay, more complex contract work to lower pay, more repetitive and strongly decomposed micro-task work.

Over the past ten years an extensive number of online marketplaces have emerged (such as Amazon Mechanical Turk; Upwork; freelancer; Innocentive; 99designs; Chaordix; OneSpace). Each online labor platforms specializes on a rather distinct set of tasks which is why the level of complexity and the degree of task-decomposition differs strongly between the platforms. On the upper end Upwork accommodates more complex tasks that are mostly not decomposed at all:

- web development (creating webpage, improving current internet pages, coding automated forms, etc.)
- IT (restructuring database, creating a failover server, migrating data, etc.)
- writing (writing a product review, writing text for homepage, writing advertisement text, etc.)
- design-tasks (designing a presentation, designing a logo, making a composite photograph, etc.)

for which workers, depending on the task, receive a compensation between $10 - $20/hour, while completing the tasks takes several hours (Agrawal et al. 2015). On the lower end MTurk is representative for highly decomposed, repetitive and less complex micro-tasks such as

- collection of information from pictures (count people on a picture, identify which groups are represented in pictures, analyze which colors are mostly used in a picture)
- providing key-words (tag pictures with key-words)
• administrative support (answering and directing phone calls, planning meetings, data entry, etc.)
• marketing (cold calling, event branding, promoting, etc.)

where workers earn an average salary of only $2/hour (Hitlin 2016), partly due to the fact that tasks are decomposed to units that small, that in most cases they pay less than $0.10/task (Hitlin 2016). Figure 4 visualizes the design space of online labor considering complexity and payment including a mapping of the two representative platforms.

![Figure 4: Mapping complexity vs. payment in a crowd work design space](image)

Reviewing this background, the question arises, whether micro-task workers, such as the crowd on MTurk may be enabled to perform more complex work. This development would release requesters from the necessity to decompose work in micro-granulated tasks and help avoid concerns going along with it and at the same time give workers the possibility to focus on solving a task and making money instead of constantly seeking new micro-work (the difference in tasks and payment is presented in Figure 5 and Figure 6 which show real examples from both platforms, MTurk and Upwork). Furthermore,
considering the poor quality of work on micro-task platforms that requesters criticize, I am interested in exploring methods that may enhance the performance of these workers.

Research presents several approaches to address this issue, focusing on either learning or improving work quality. These approaches are explained in the following chapter.

<table>
<thead>
<tr>
<th>Classify and mark the differences between 2 pages</th>
<th>HIT Exploration Date: Jul 3, 2017 (1 day 1 hour)</th>
<th>Reward: 0.08</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description: Compare two pages and find difference</td>
<td>Time Allocated: 20 minutes</td>
<td></td>
</tr>
<tr>
<td>Keywords: article, data, details, announcement, people, USA, official, scam, compare, two pages, find difference</td>
<td>Qualifications Required:</td>
<td></td>
</tr>
<tr>
<td>Total approved HITs is not less than 5000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIT approval rate (%) is not less than 98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MBA Announcement is not less than 90</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Count the words, images and videos in the page</th>
<th>HIT Exploration Date: Aug 31, 2017 (3 weeks 3 days)</th>
<th>Reward: 0.08</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description: Answer the questions of counting the number of words, images and videos in the given page link</td>
<td>Time Allocated: 15 minutes</td>
<td></td>
</tr>
<tr>
<td>Keywords: words, images, content extraction, duration</td>
<td>Qualifications Required:</td>
<td></td>
</tr>
<tr>
<td>Total approved HITs is greater than 100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HIT approval rate (%) is greater than 98</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5: Real work examples from MTurk

Figure 6: Real work examples from Upwork

2.2. Learning

2.2.1. Formal Education

One solution is that workers could seek to gain knowledge and expend their skillset in more formal education settings. Workers can enroll in all levels of schooling or vocational
programs. With the development of the internet and mobile technologies, additional options to advance have become available. In massive open online courses (MOOCs) students can participate remotely in activities (i.e., lectures, tutorials, and assignments) intended to simulate a “real” learning environment without the need of being physically present at a teaching facility. While some MOOCs charge a fee, many offer free courses (Coursera). Further, a number of educational environments have created applications for mobile devices—such as Duolingo for language learning (Duolingo), SoloLearn for programming (SoloLearn) and Khan Academy for subjects like math and physics (Khan Academy)—allowing learners to access exercises for short periods of time while on the go.

To go beyond the traditional environment of formal education there has been much interest in exploring the possibility of authentic learning, so learners can apply the knowledge acquired in the classroom. Various fields such as instructional design, media and communication, history, hospitality education were subject of studies for designing authentic tasks (Clinton & Rieber 2010; Collis, Foth & Schroeter 2009; Deale, Elders & Jacques 2010; Herrington 2006). A variety of different practical applications such as listed in the following realize the concept of authentic learning. Gerber, Marie Olson & Komarek (2012) introduce Extracurricular Design-Based Learning (EDBL) which is a student initiated and student directed learning. Interdisciplinary student teams identify actual, pro social and local community challenges and apply the human-centered design and innovation process to develop and implement creative solutions. Online learning environments such as Mekong e-Sim simulate role playing to train participants in complex authentic decision making (McLaughlan & Kirkpatrick 2004). Applications such as working with remote instruments operated through a browser interface where students around the globe can conduct experiments with specialized equipment were developed at MIT’s labs (Lombardi 2007).

Acquiring knowledge is certainly an effective and sustainable long-term solution providing a foundation for overcoming knowledge-based work limitations, but it requires intense time investment and dedication for a considerable period of time until a certain level of proficiency can be achieved. Hence, it may not be applicable for all members of
the community that in average take on work on online platforms that pays under minimum wages (according to the minimum wages laws in the US) (United States Department of Labor 2017; Boonthum-Denecke 2011; Ipeirotis 2010b; Ross et al. 2010; Agrawal et al. 2015).

2.2.2. Learning from Feedback in Crowdsourcing

As learning requires both practice and feedback (Sadler 1989) a number of commercial as well as academic efforts have focused on exploring the use of feedback for learning purposes in the crowd environment.

Many online learning environments have built systems for providing students automatic feedback on quizzes and homework. Automated feedback was applied in the context of essay grading and design (Hearst 2000; Fischer et al. 1993). Chen (2004) developed a system which automatically analyses students’ submitted programs in a computer science course and gives immediate feedback. Automated feedback based techniques have been developed for the discovery of better student models (Koedinger, McLaughlin & Stamper 2012).

As presented, automated feedback in a crowd work context could enhance learning and lead to better results. Unfortunately, automated approaches only work when the problems have known solutions, not for more creative and complex task. Requesters on online platforms hire workers to complete tasks, which they would not do if they were already in possession of a solution. Hence, automated feedback is not compatible with the idea and purpose of paid crowd work.

To extend the effect of learning through feedback beyond automated approaches, systems like PeerStudio (Kulkarni, Bernstein & Klemmer 2015) have been investigating how to scale peer feedback for more open-ended project-based courses like interaction design. However, to achieve improvement of work and learning effects this approach requires continuous exchange between peers, which may not be applicable in an online work environment.
2.3. Improving Work Quality

To improve the quality of online work, scholars have researched approaches such as coordination and feedback.

Coordination intends to guide workers and tunnel efforts in order to avoid redundancy and to lead the crowd efficiently towards a common goal (Kittur & Kraut 2008; Kittur, Chi & Suh 2008). It requires a leader or a group that embodies an organizational unit constantly supervising and structuring work and its progress. In the particular case of crowd work, this could be realized by either the requester or dedicated resources provided by the platform.

While organizations outsource tasks to save costs and improve their efficiency, they may not be interested in spending additional time and money on supervision. Online labor platforms on the other hand do not experience any direct benefit from helping improve the work quality, beside a potential improvement of their image. This may not be enough of an incentive to continuously dedicate time and money resources to this topic.

Feedback as a form of communication was proven to lead to significantly better results. Zhu et al. (2014) discovered that giving feedback on other peers’ work improves the quality of one’s own subsequent tasks. Furthermore, feedback from interactive reviewer teams was proven to outperform individual feedback. Dow et al. showed that timely and task-specific feedback from the requester helps to shepherd the crowd and improves work quality (Dow et al. 2012). On platforms such as Forrst (ZURB 2015), Photosig (Xu & Bailey 2012) and Dribble (Marlow & Dabbish 2014) users can provide peer-based feedback. Luther et al. showed that crowd feedback exchange improves when workers use structured rubrics that outline important criteria (Luther et al. 2015). Other studies such as the micro-task platform Atelier have explored the use of hired professionals for mentoring and providing feedback for real crowd work tasks (Suzuki et al. 2016). Drapeau et al. (2016) developed MicroTalk, a workflow to improve quality of crowd work based on feedback in form of self-assessment, justification of own work and reconsideration due to peer feedback. While these feedback methods have proven
valuable, the downside is that they require additional human resources, leading to the same scenario as the approaches explained before.

2.4. Applying Scalable Methods to Crowds

Unfortunately, both traditional educational paths and the previous attempts to improve work quality require time investment and commitment by learners and/or requesters. Inspired by the theory of “zone of proximal development” (Vygotsky 1980) that explains the difference of what novices can do with additional help, this research investigates the use of scaffolding (van der Stuyf 2002; Hammond und Gibbons 2005; Reiser 2004; Hogan und Pressley 1997) for online work to enable micro-task workers to perform more complex tasks and achieve better quality of work. In this thesis I explore if this can be achieved by giving appropriate guidance applying four (a priori determined) proven methods from the learning science literature: examples, rubrics, motivating task rationale and step-by-step instructions.

This approach has been chosen as subject for my thesis based on its potential to be applicable with minimal additional effort. The only extra effort that would go along with it is the time investment of requesters putting a task online who would have to write their task description more detailed, giving additional guidance.

In the following the proposed four methods for scaffolding are explained in detail.

2.4.1. Examples

Examples provide a powerful tool for learning domain-specific principles (Colhoun, Gentner & Loewenstein 2008; Gentner, Loewenstein & Thompson 2003; Lee et al. 2010; Thompson, Gentner & Loewenstein 2000). They illustrate a pattern that may be repeated and applied when solving a similar task. By combining explanations with the possibility of comparison, synergies can be created (Williams, Lombozo & Rehder 2013) and even greater learning benefits may be yielded beyond engaging in just one of these processes (Edwards et al. 2014). Hence, reviewing different examples allows one to recognize similarities as well as differences between the cases, prompting category learning and abstract relational schemas which eventually leads to a better understanding of the task
(Kurtz, Boukrina & Gentner 2013; Edwards et al. 2014). Kulkarni, Dow and Klemmer showed in an online creativity experiment that early exposure to examples improves the outcome. Even better results were yielded by repeated exposure to examples. To meet the need for examples researcher have been working on developing exploratory techniques for finding relevant and inspiring examples (Ritchie, Kejriwal & Klemmer 2011).

2.4.2. Motivating Task Rationale
Completing meaningful tasks leads to motivation at the workplace (Rosso, Dekas & Wrzesniewski 2010). Meaningful in this context implies that the worker is both doing work with purpose and receiving acknowledgements for accomplishments (Ariely, Kamenica & Prelec 2008). Chandler and Kapelner (Chandler & Kapelner 2013) transferred these findings into the crowd environment showing an interdependency between how a task is framed and outcome in terms of work output. Motivating task rationale in terms of expressing a purpose and higher goal led to a significantly higher willingness for participation and quantity of output.

2.4.3. Rubrics
Rubrics are a set of criteria or dimensions of quality for a given type of performance. Using rubrics can help to formulate standards for accomplishment and thereby guide as well as improve performance (Allen & Tanner 2006). Furthermore, they can serve as a medium to capture and articulate specific standards clearly and explicit (Allen & Tanner 2006). Rubrics can be applied for a large variety of tasks (Stevens & Levi 2013). In crowd context using rubrics has already been widely researched. Systems such as Shepheard (Dow et al. 2012) showed that rubrics enhance novices to provide feedback at an equal level with professionals (Yuan et al. 2016). PeerStudio (Kulkarni, Bernstein & Klemmer 2015) explored the possibilities of using rubrics for peer-based feedback in a MOOC setting. CrowdCrit (Luther et al. 2015) leveraged paid crowdsourcing to generate high-quality critique for visual design based on rubrics providing a collection of design principles and critique statements.
2.4.4. Step-by-step Instructions

Decomposing tasks between workers in the context of crowd work has been widely researched and applied. Many scientific contributions deal with creative methods exploring how tasks can be decomposed (Kittur et al. 2013; Bernstein et al. 2015; Teevan et al. 2016; Nebeling et al. 2016). Requesters on online labor platforms are showing a trend towards increasingly smaller work units they offer (Teevan, Liebling & Lasecki 2014; Teevan et al. 2016). Decomposing tasks given to a single worker has on the other hand not found much interest or application in the crowd environment. Though, established teaching theory proves that breaking tasks down into small, manageable units is the best approach to communicate knowledge and hence the most effective way to enhance learning (Mergel 1998). Step-by-step instructions provide a sequential list of operations and still allow a general overview of the complete task, which was shown to yield better results (Lasecki et al. 2015; Cai, Iqbal & Teevan 2016).
3. Own Approach

In order to answer the stated research questions this study follows a between-subjects design exploring the factor “task description” with the two conditions original versus scaffolded. The data was collected on two tasks from different genres (product review writing and slide design). The tasks were extracted from the online labor platform Upwork. The original condition uses the task description found on Upwork without modifications. The scaffolded condition reframed the instructions using the template of learning techniques proposed in 2.4. Participants from MTurk performed the task under either original or scaffolded condition. Participants from Upwork performed tasks under original conditions. The submitted results were compared based on a pre-defined set of criteria. The setup of the study is visualized in Figure 7.
3.1. Participants

The goal of the study is to investigate if micro-task crowd workers can be enabled to achieve better results and perform more complex online work tasks (delivering outcome
comparable to the level of performance of professionals active on platforms hosting more complex tasks). Therefore, the participants were recruited from two distinct pools – a micro-task and a complex-work environment. Since MTurk is the best-known low-end micro task platform 114 workers were recruited from MTurk as representative participants for the target group of this study that shall be enabled to perform at a higher level of quality and more complex tasks. To validate if the work of MTurk workers can achieve a similar level of quality as work performed on high end platforms, additional 6 workers from the upper-end platform Upwork were recruited as a reference for comparison. Unfortunately, due to the limited budget I had for my research, I was not able conduct my study with bigger sample sizes. Because work on Upwork is paid significantly higher fees I was only able to afford recruiting 6 Upwork workers for my study.

This study is focused on the US labor market, hence only American participants were allowed. Of the 114 participants recruited on MTurk 34 were female and 80 male with an average age of 30.5 years. The participants from Upwork were 2 females and 4 males with an average age of 24.2.

3.2. Subject of Study – the Tasks

To investigate whether micro-task workers can be enabled to perform more complex, higher-pay tasks the study is conducted based on representative tasks from the upper end crowd work platform Upwork. The criteria that the choice of tasks was based on are:

- Occurrence – Tasks with high rate of occurrence on the upper-end platform Upwork
- Diversity - Tasks from different fields
- Pre-existing knowledge – Tasks with different required level of pre-existing knowledge, however not “too” specific to not limit the pool of potential participants
- Difficulty level - entry level or intermediate skills so novices can solve them
Based on Agrawal et al (Agrawal et al. 2015) extensive study that analyzed tasks and payment on o’Desk\(^1\) as well as own investigation a writing task and a slide design task were chosen as appropriate and representative tasks for the purpose of this study. For both tasks, a version under the original as well as scaffolded condition were created. In order to avoid confusion of participants, to make the results more comparable and to enable their completion within reasonable time (max. 20 min per task) minor modifications in terms of scope and wording were made to the original wording of the task description under original condition. The scaffolded versions of the tasks apply the previously proposed teaching techniques for framing the task: examples, rubrics, motivating task rationale and step-by-step instructions.

3.2.1. Scaffolding

There are various ways how to apply the proposed teaching techniques in the context of scaffolding. Since this study is aiming at finding a methodology that requires as little additional investment of time, money and human resources as possible, to provide best possible guidance for workers, scaffolding was realized as explained in the following:

Examples

To give the learners a general idea of how the specific task could be solved they were provided multiple prototypical examples from a different subject. The writing task description contained for each section of a product review (introduction, main body, verdict) several examples. The examples for the introduction are presented below:

“For a while now I’ve wanted to switch to an all natural shampoo, because the sulphates and other chemicals have been making my scalp itchy”
“I’ve been a fan of all of their other rose products — but this one really worked for me — and with a quickness. I started using this last Friday night”

---

\(^{1}\) The study conducted by Agrawal et al. (2015) investigated payment levels for different types of tasks on the online labor platform o’Desk. This platform does not exist as a separate unit since it merged with eLance in 2013 to Upwork. Because Upwork is one of biggest online freelance market and o’Desk a significant part of it, this study is uses the findings of Agrawal et al. to identify representative tasks.
The examples for the slide-design task showed several presentation slides that are designed according to the relevant design standards1 (see Figure 8) and an example how the slides should not be designed.

![Figure 8: Example of a well-designed slide](image)

Motivating Task Rationale
Explaining the purpose of work and putting it in a bigger picture is proven to lead to higher motivation and better results of work. Hence, both scaffolded task descriptions motivate the task rationale. The writing task was motivated as follows:

“To help other customers determine which mobile phone suits their needs best and may be the right choice for them we would like to gather product reviews of mobile phones. Therefore, we would like you to write a review of a mobile phone of your choice (you may review your own phone).”

1 Design and is always subject to a subjective experience and leaves room for personal interpretation and evaluation. What some people find to be well-designed others may find confusing or inappropriate. However, in the field of design researchers have identified prevailing key principles and factors that are generally applicable leading to a well-designed result: Readability, Layout, Balance, Simplicity, Emphasis, Consistency and Appropriateness Nielsen & Molich (1990); Wattenberg & Kriss (2006). These factors are visualized in the examples for the slide-design task and later on used for evaluating the learners’ submissions.
Rubrics
Learners are often inexperienced and do not yet fully understand what is crucial to solve a task successfully. Hence, they were provided a list of criteria specific for each task, which explains what their solution should include and what they should make sure to consider when solving the task. The rubrics provided for the writing task were:

- it is specialized for the customer group considering all information they would be interested in
- it contains a balanced evaluation of all important aspects
- it comes to a clear conclusion
- it is catchy and if it sounds convincing and trustworthy
- the grammar and spelling is correct

Step-by-step Instructions
Without prior knowledge of how to structure a workflow in order to achieve a good result novices may, especially when working under time pressure, submit a result that does not present their actual abilities. The scaffolded task descriptions contain therefore a task description that breaks the task down into smaller units that students can solve sequentially. The writing task as well as the slide-design task were divided into three subsections:

1. Before the learner starts working on the task
2. Solving the task
3. Reviewing the solved task

The writing task realized this as follows:

“**Before** you start writing your review, you should prepare yourself and make sure that you:

*Know the product:* ...
*Know the need it meets:* ...
*Know the customer:* ...

**Writing** the review follow this structure:

1. Introduction
2. Main Body
3. Verdict

**Before you finalize** your work revise your review checking if:
3.2.2. Writing task

A common task on Upwork is to write a review. After reviewing multiple possibilities, I selected the following task from Upwork (original wording) as subject of the study due to its good fit with the criteria explained in 3.2:

“I am looking for Product Review Writer, including children's clothes, toys, skin care products, women and family products. All articles must be 100% unique, have excellent grammar and outstanding writing style. I prefer you can write from a mom's point of view and in a blogger style. Minimum words is 500.

Writer must be able to perform detailed research about the subject area. You will have to look up product description, customer reviews at other websites and create a good flowing and personal product review.

Articles must pass copyscape and duplchecker otherwise they will not get paid for.”

The writing task under original condition is presented in Figure 9:

“I am looking for a Product Review Writer for a mobile phone. Please, write a review of a mobile phone of your choice. All articles must be 100% unique, have excellent grammar and outstanding writing style. I prefer you can write in a blogger style. … Articles must pass copyscape and duplchecker otherwise they will not get paid for.”

Figure 9: The writing task under original condition
The writing task under scaffolded condition can be found in Appendix A. Figure 10 presents an extract of the scaffolded task emphasizing the implementation of the four teaching techniques.
To help other customers determine which mobile phone suits their needs best and may be the right choice for them, we would like to gather product reviews of mobile phones. Therefore, we would like you to write a review of a mobile phone of your choice (you may review your own phone)....

**Before** you start writing your review, you should prepare yourself and make sure that you:
- Know the product: Be familiar with the product, it makes it easier to appear authentic and trustworthy.
- Know the need it meets: ....

**Writing** the review follow this structure:

1. **Introduction**: Introduce the product as well as the promise it makes.
   - Engage the reader to continue reading
   
   **Examples:**
   “For a while now I've wanted to switch to an all natural shampoo, because the sulphates and other chemicals have been making my scalp itchy” ...

2. **Main body**: Introduce the specific product, try to provide a descriptive visualization of the product, add practical details (price, size, weight, etc.), you may as well include a comparison to other products
   - Be a specific and detailed as possible
   
   **Examples:**
   descriptive visualization: “It smells good (although a little strong- not a kind to use in public) and is thick enough to soak deeply into my skin so I don't need to reapply. I do find it a little too thick to use on my hands or when it's just a normal day” ...

3. **Verdict**: Provide a balanced evaluation of pro's and con's and an estimation either the product delivers its promise and summarize your statement
   - Be convincing without lobbying
   
   **Examples:**
   “I liked that it's tinted, and self tans. The self tanning lasts a few days, as it's very subtle, not the 2 weeks it claims. But the bronzing is immediate.” ...

**Before you finalize your work revise your review checking if:****
- it is specialized for the customer group considering all information they would be interested in
- it contains a balanced evaluation of all important aspects  ...

---

*Figure 10: Extract of the scaffolded writing task.*

*Note: The applied teaching techniques are marked as following: light blue = Motivating task rationale; dark blue = Step-by-step Instructions; light gray = Examples; dark gray = Rubrics*
3.2.3. Slide Design Task

The second task chosen for this study was a slide design task. Several tasks on Upwork were reviewed. Unfortunately, none of the tasks found was appropriate because of the visual material they provided which the presentation was supposed to be based on. Hence, for this study a slide-design task was designed, that is representative for the slide-design tasks that can be found on Upwork, however, its content does not violate any privacy policies and it does not interfere with the actual work tasks on the platform.

The slide-design task under original condition is shown in Figure 11:

I need a 2-slide presentation on my company summarizing who we are and what we do. I will provide you with a link to pictures and a summary of information that the presentation will be based on, you will have to put the relevant information into slides for our audience.

Link: [https://moocita.org/](https://moocita.org/)

Content: We are a Non-Profit organization. We help our students to acquire job relevant skills by providing free online courses and assignments that are real world job tasks. Students solve these tasks alone or as teams and get paid if the quality of their work meets the necessary requirements. Working on real world tasks helps students to develop their portfolio and keep track of their progress. Instead of grades students see an actual impact on their bank account and receive letters of recommendation from employers.

Figure 11: Slide-design task under original condition

The slide-design task under scaffolded condition can be found in Appendix A. Figure 12 presents an extract of the scaffolded task emphasizing the implementation of the four teaching techniques.
We would like your support in creating a 2-slide presentation of our company that shows: Who we are & What we do

Below you find a summary of the content we would like to be included as well as a link to our homepage providing you pictures of our team.

**Link:** [https://inocita.org/](https://inocita.org/)

**Content:** We are a Non-Profit organization. We help our students….

When creating a powerful presentation, you should follow the guidelines listed below:

1. Understand the aim of the presentation and who the audience is
2. Once you start working on the presentation, make sure to include these rubrics:
   - **Readability:** No spelling, grammar or punctuation mistakes, …
   - **Layout:** Fonts: fit the purpose… Colors: support/highlight the content…
   - **Balance:** Symmetry is used to create a sense of order/movement/change…
   - **Simplicity:** Not too many information, bullet points instead of sentences…
   - **Emphasis:** Visual proximity to related content, no overemphasis…
   - **Consistency:** Organizational scheme, visual cohesiveness
   - ** Appropriateness**
   - The design successfully communicates a single clear message, …

3. Before you submit your work, check again if you followed all rubrics

Here you find examples of “good” and “bad” slides:

![Example slide 1](image1.png)

“bad slide”

![Example slide 2](image2.png)

“good slide”

**Figure 12:** Extract of the scaffolded slide-design task.

*Note: The applied teaching techniques are marked as following: light blue = Motivating task rationale; dark blue = Step-by-step Instructions; light gray = Examples; dark gray = Rubrics*
3.3. Procedure

The experiment consisted of 2 main phases - data collection and evaluation of submissions.

3.3.1. Data Collection

The first phase of this experiment was the collection of data. For gathering data on novice work the study was carried out first at MTurk. The participants on MTurk applied for one of the two tasks this study is based on. They were randomly assigned a task description in original or scaffolded condition of the respective task. The participants working under the original condition were serving as the control group. The participants on MTurk were chosen following the first come first serve principle. To avoid distortion of the results by unequal language proficiency the participants of the study were intended to be living in the United States. Because MTurk does not allow to limit the pool of potential workers to a specific geographic region, every participant had to confirm his nationality as part of the consent form, filled out before the task was handed out. To successfully complete the task the workers had to submit their result and fill out a questionnaire providing basic demographics.

Then, to have a professional reference to compare the work of MTurk novices with, workers from the platform Upwork were recruited. To guarantee the workers chosen were representative for the high level of work quality offered on Upwork, only applicants with a ranking of 4 and more stars (out of 5) were considered. Under this restriction, again, the approach first come first serve was applied. The assigned workers on Upwork were handed out the respective task under original condition to provide authentic insight on the actual level of quality of work, when Upwork workers solve tasks on the platform. Again, to successfully complete the task the workers had to submit their result and fill out a questionnaire providing basic demographics.

3.3.2. Evaluation of Submissions

In this phase, the submissions from both platforms were evaluated. Each of the tasks had a different set of evaluation criteria.
The criteria of the writing task are listed in Figure 13. They represent a summary of best practice, guidance from blogging sites and demands requesters expressed in the writing task they posted on Upwork, which I generated based on my research to this topic.

**Figure 13: List of criteria for evaluating the writing task**

- The worker wrote an original review (did not plagiarize)
- The worker had no spelling, grammar or punctuation mistakes
- The worker included different sections (introduction, main body, verdict)
- The worker wrote an engaging, authentic review
- The worker includes personal stories/anecdotes
- The worker included a description of the product
- The worker assessed positive and negative aspects of the product
- The worker provides a clear verdict

The criteria the evaluation of the slide-design-task was based on are presented in Figure 14. This is a template provided by researchers who performed a bottom up analysis of relevant sources concerning best practices and principles. Inspired by previous efforts of the HCI community (Nielsen & Molich 1990; Wattenberg & Kriss 2006) they identify the following seven key design principles: Readability, Layout, Balance, Simplicity, Emphasis, Consistency and Appropriateness (Luther et al. 2015).

For each task the list comprised 8 features for successful completion. The sum of checked features provided the experts’ evaluation of task performance. The final Task Performance was calculated as the average of the experts’ grading.

The submissions were reviewed by external academic professionals specialized within the area of product and service design. From a pool of potential judges these four professionals were chosen based on their knowledge within this field and because of their not existing pre-knowledge of this study. Each task was reviewed by two judges to minimize the effect of subjective preferences on the results. The evaluation was conducted double-blind to avoid bias by having access to information that may influence the behavior of the judges.
3.3.3. Payment

The wages paid were considered carefully. Given that MTurk and Upwork offer different payment models and market rates, workers on MTurk were compensated with $2.50/task and workers on Upwork with $10/task. These numbers ensure alignment with the market, whereby the introduction of additional confounding variables was avoided.

3.4. Measures

To investigate the stated research questions this study analyzed the effect of three independent variables (task description, task, and population) on the dependent variable work quality.
3.4.1. **Independent Variables**

The independent variable *description* of a task has two levels: scaffolded indicates the described scaffolding method was used, while original indicates unaltered task descriptions.

The variable *task* indicates which of the two tasks participants solved. The variable has two levels, either designing slides (slides) or writing product descriptions (writing).

The variable *population* encodes from which online labor market participants were recruited, either MTurk or Upwork.

3.4.2. **Dependent Variables**

The dependent Variables investigated in this study were Task Performance, Satisfaction with Task Description, Satisfaction with own performance and Most Helpful when Solving the Task.

*Task Performance:* The task performance is determined by the individual evaluation of each submission. Per task two independent referees (in total four – two per task), blind to condition using a check list (specific for the respective task) assessed which criteria were met. The final Task Performance of each submission was calculated as the average of the experts’ grading, on a scale from 0-8.

*Satisfaction with own performance:* Furthermore, I collected data on the self-assessment of the novice workers. Directly after submitting their work, novice workers judged their Satisfaction with own performance on a scale from 0 to 10 (0 = not satisfied; 10 = very satisfied). The scaffolded task descriptions were designed applying a template of several teaching techniques.

*Most helpful when solving the task:* To understand the novices’ perspectives on the techniques, I provided scaffolded participants an open question asking what they found most helpful when solving the task.
4. Evaluation

This chapter contains the analysis of the data set collected during the experiment as well as the interpretation of the results. The different statistical methodologies used for the data evaluation are explained in detail. Furthermore, conclusions are deduced from the evaluation’s results.

4.1. Data Analysis

First, main and interaction effects were analyzed. To investigate these a linear model was used and an analysis of variance (ANOVA) was applied. To ensure that the data set meets the assumption of the ANOVA test a Bartlett test of homogeneity of variance was performed for all grouping parameters. The Bartlett test of homogeneity showed all p values being above 0.35. Henceforth, homogeneity for my data set within acceptable ranges was assumed. Testing the \( H_0 \)-hypothesis “The independent variable D/P/T does not have any influence on the dependent variable task performance” the ANOVA showed significant results for all three main effects. The \( H_0 \)-hypothesis was rejected based on the significantly low p-values. The test for interaction effects between any levels of the main variables showed no interaction effects. Both results are presented in table 1. (The description variable (D) indicates if scaffolding was used. The population variable (P) encodes if the participant was recruited through Upwork or MTurk. The task variable (T) indicates which task (writing, slides) was solved.)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F</th>
<th>p</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(D)escription</td>
<td>1</td>
<td>26.5</td>
<td>26.52</td>
<td>6.227</td>
<td>0.014</td>
<td>*</td>
</tr>
<tr>
<td>(P)opulation</td>
<td>1</td>
<td>32.6</td>
<td>32.59</td>
<td>7.653</td>
<td>0.006</td>
<td>**</td>
</tr>
<tr>
<td>(T)ask</td>
<td>1</td>
<td>61.4</td>
<td>61.47</td>
<td>14.43</td>
<td>0.0002</td>
<td>***</td>
</tr>
<tr>
<td>D x P</td>
<td>1</td>
<td>0.97</td>
<td>0.971</td>
<td>0.228</td>
<td>0.633</td>
<td></td>
</tr>
<tr>
<td>P x T</td>
<td>1</td>
<td>5.05</td>
<td>5.054</td>
<td>1.186</td>
<td>0.278</td>
<td></td>
</tr>
<tr>
<td>Residuals</td>
<td>114</td>
<td>485.53</td>
<td>4.259</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Main and interaction effects between the three independent variables.
Second, Table 2 shows the exact performance evaluation of MTurk workers under both conditions as well as the work from Upwork.

<table>
<thead>
<tr>
<th>Population</th>
<th>Description</th>
<th>Task</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTurk</td>
<td>scaffold</td>
<td>writing</td>
<td>6.06</td>
<td>1.9</td>
</tr>
<tr>
<td>MTurk</td>
<td>original</td>
<td>writing</td>
<td>5.01</td>
<td>2.1</td>
</tr>
<tr>
<td>Upwork</td>
<td>original</td>
<td>writing</td>
<td>5.00</td>
<td>0.5</td>
</tr>
<tr>
<td>MTurk</td>
<td>scaffold</td>
<td>slides</td>
<td>4.54</td>
<td>2.7</td>
</tr>
<tr>
<td>MTurk</td>
<td>original</td>
<td>slides</td>
<td>3.61</td>
<td>2.0</td>
</tr>
<tr>
<td>Upwork</td>
<td>original</td>
<td>slides</td>
<td>1.50</td>
<td>0.8</td>
</tr>
</tbody>
</table>

Table 2: Summary of work performance in all six possible settings

To investigate the direction of effects a multiple linear regression was used. As the interaction effects were not significant only the main effects were analyzed to prevent inflating type II errors. Comparing the cumulated results of all 114 MTurk participants of which 43 solved the writing and 71 the slide design task in either the scaffolded or the original version, the scaffolded versions (row Description Scaffold) received an average final rating higher by one point than the original versions. Work samples collected through Upwork are on average of lower quality than the results collected from MTurk given the scaffolding method was used. A summary of the multiple linear regression model is shown in table 3.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>t-value</th>
<th>p-value</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>6.074</td>
<td>0.301</td>
<td>20.18</td>
<td>&lt;0.001</td>
<td>***</td>
</tr>
<tr>
<td>Description</td>
<td>1.001</td>
<td>0.391</td>
<td>2.563</td>
<td>0.011</td>
<td>*</td>
</tr>
<tr>
<td>Scaffold</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Population</td>
<td>-2.038</td>
<td>0.895</td>
<td>-2.282</td>
<td>0.024</td>
<td>*</td>
</tr>
<tr>
<td>Upwork</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: Summary of the multi-linear regression model
When performing a $\chi^2$ test, Mturk workers using scaffolding outperformed participants from Upwork in both tasks. The direct pairwise comparison showed, that experts chose scaffolded work from MTurk significantly more often than work from UpWork (77.8% vs. 22.2%). MTurk workers stated examples and step-by-step instructions to be most helpful to successfully solve the task. However, the self-assessment survey showed no significant difference in satisfaction with work performance (scaffolded M = 6.92 vs. original M= 7.63) or satisfaction with task description (scaffolded M = 7.58 vs. original M= 7.34).

To ensure that the rating scheme is viable and not significantly dependent on the individual opinion of a rater, the Pearson Product correlation between the two independent raters was calculated. The raters opinion can be assumed to be objective since the test shows that the rater who evaluated the results independently agreed almost perfectly on the rated workpieces $r = 0.91$, $t(119)=23.33$, $p < 0.001$.

The workers claimed examples and step-by-step instruction to be the scaffolding methods they valued most when solving tasks und scaffolded condition.

4.2. **Data Interpretation**

4.2.1. **The Performance of Micro-Task Workers is Being Underestimated**

According to the sample analyzed in this study, workers on MTurk are able to perform more complex tasks as offered on professional platforms such as Upwork. In general, there was no significant difference in performance in terms of work quality ($p > 0.3$) between the Upwork population and the MTurk population in this study. This holds true for both tasks, the writing as well as the slide design task. The results in table 2 show that the original submissions from MTurk were of at least same quality as submissions from Upwork. In the writing task, workers from Upwork working under original condition received an average rating of 5.00 while MTurk workers working under same condition delivered an average performance of 5.01. The results of the slide design task were even
more significant. While Upwork workers under original condition received a rating of 1.5, workers from MTurk working on the same task received a rating of 3.61.

4.2.2. Scaffolding, Leads to Better Performance

When using scaffolding for task descriptions the performance of MTurk workers was significantly improved. MTurk workers solving the writing task under scaffolded condition received a rating that was in average 1.05 better than the rating under the original condition. The performance in the slide design task was enhanced by 0.93 when scaffolding was used.

This result is additionally confirmed by a direct pairwise comparison where judges chose in more than three out of four cases a solution MTurk workers delivered when working on a scaffolded task over a solution from Upwork workers. However, in the self-assessment survey micro-task workers did not see a significant increase in their performance (scaffolded M = 6.92 vs. original M= 7.63). The opposite was the case. The workers judged their performance to be better when working without scaffolds.

Furthermore, even though scaffolded task descriptions enabled the participants to achieve significantly better results, they did not express any significant difference in satisfaction with the task descriptions (scaffolded M = 7.58 vs. original M= 7.34).

Examples of the submissions under the different conditions are presented in Figure 15.
Figure 15: Examples of submissions - slide design task

**Mooqita**

- Our mission is helping students
  
  we provide a skill development programs
  
  Earn a living

MTurk
Condition: original

Condition: scaffolded

Upwork
Condition: original

We help students find the job they love by bringing education and work closer together.

We research the
effects of rubrics:

- We design to support students studying online education.
- We help them earn money while they do.
- Find them a job they love.

Investigate a design to enable MOOC learners to apply their skills and earn money in an online marketplace.
5. Discussion

The following chapter reflects upon the results of the study. In the first section the results are summarized followed by a discussion of these in section two. Section three points out possible implications. Section four points out the strength and weaknesses of the study.

5.1. Results of the Study

The research I have done within this thesis leads to the following results:

1. In my study workers using scaffolding scored significantly higher results (over 10%) than workers that were not provided scaffolded task descriptions.
2. Workers from the lower-end platform MTurk performed at least on par with professionals. In the slide design task, they even outperformed workers from Upwork receiving ratings that were in average over twice as high.
3. Performing a $\chi^2$ test the scaffolded MTurk submissions were preferred in 77.8% cases over the original Upwork conditions.
4. Workers were more satisfied with their own performance when working under the original condition than under the scaffolded one (scaffolded = 6.92; original=7.63).
5. The workers were slightly more satisfied with the scaffolded task descriptions than with the original ones (scaffolded = 7.58; original = 7.34).
6. Workers found examples and step-by-step instructions to be most helpful when solving tasks.

5.2. Discussion of the Findings

Some of the results were expected and proven by the study I conducted. This was the case with finding number 1. In traditional learning, teaching techniques help novices produce better results (Chandler and Kapelner 2013; Yuan et al. 2016; Kurtz, Boukrina and Gentner 2013; Mergel 1998). I expected to observe the same effect when applied on online tasks and the results of my study validated my expectations as scaffolding lead to significantly higher scores. This result aimed at answering the second research question which asks, if task descriptions enhanced by using scaffolding enable novices to achieve
higher quality of work, but I cannot make a general statement that this holds true. I do not think that when speaking about tasks that do not have just one right answer, but which strongly depend on the subjective preferences of the judge the general term “quality” can be used. The evaluation of my study was done by independent judges, based on a set of criteria, however, the list of criteria was generated by me. I did not include any aspects that I personally would have wanted to be represented, but even though I tried to gather an objective summary of best practice, I still decided based on my personal view, which aspects should be included or left out. Therefore, the list of criteria used for the evaluation represents a subjective foundation for evaluation and should not be used as judgment for general quality. What evaluating the submissions based on my lists of criteria actually showed was how satisfied the requester (in the study me) was with the submissions based on the task description the requester (me) generated and how scaffolding according to the requesters (mine) expectations could influence the satisfaction with the submissions received. This interpretation of the term “quality” and the fact that it cannot be generalized does not compromise the results of my study, though. In the end, what it comes down to on online labor markets is the opinion of the requester who is posting and paying for a task. This requester must be satisfied with the results, his/her personal task must be represented in the results, not any general metrics which may objectively classify the quality of the received work. This is a perspective which the study conducted in my thesis mirrors. Hence, the second research question can be answered with “yes, scaffolding does enable novices to achieve higher quality of work, if quality is defined in terms of satisfying the requester by meeting his/her expectations”. This statement is additionally supported by findings number 3 which show that judges preferred, based on their very personal taste, scaffolded work in 77.8% of the cases.

A reasonable objection which should be discussed in the context of these two research questions is the limit of scaffolding as means to enhance satisfying performance. First, I would like to define what, in my opinion and in line with the results of this study, scaffolding can and cannot do. This study has shown that scaffolding can enable workers to perform work which requesters find more satisfying. I dare to claim that independently of a task’s complexity scaffolding, which communicates expectations clearer and gives
guidance on how to achieve the result the requester is aiming at, has this positive effect on the outcome. What scaffolding did not do was to enable workers to perform tasks for which they did not have the basic skillset for. While some workers may not have written a product review before, they did have the knowledge of how to write and they may as well have seen a product review before. In this sense scaffolding enabled workers to apply their skills in a new context, but it did not generate basic skills. If somebody does not know what power point is, a scaffolded task description will not lead to any better slides because the basic skills needed are missing. Therefore, I doubt that scaffolding can enable everyone to perform better in every task presented. What I do believe scaffolding can do is, as shown in this study, to support workers to deliver more satisfying results when performing tasks that they do have a basic understanding of.

The most surprising and provocative finding of this study was finding number 2 which shows that the tasks used in this study could not just be solved by novices when scaffolded. Scaffolding allowed micro-task workers to achieve significantly better results than without but even without scaffolds, when provided the exact same task descriptions micro-task workers were able to perform at least on par with professionals from Upwork. In the slide-design task they even outperformed the professionals by scoring results twice as high. This makes the suspicion arise that the abilities of micro-task workers are underestimated. A valid objection is that the tasks which were subject of the analysis in this study are not tasks of a high level of complexity, which one could suspect, is why micro-task workers are able to perform equally well as professionals. This statement however, does not take into consideration what tasks micro-task workers normally work on. Task on upper-end platforms are repetitive tasks that only require basic cognitive skills, tasks such as counting the number of colors used on a web page for example. A task where workers are asked to design a slide requires a significantly broader and more developed set of skills. Workers must extract relevant information, insert them in an appropriate form in the slide, create a design, decide how to communicate the messages, and much more. Therefore, I dare to state that abilities of micro-task workers on lower-end platforms are underestimated. However, this statement should not be confused with the claim that workers on Upwork and MTurk have the exact same set of skills. I strongly
doubt this as I do not believe that somebody who can solve quite complex tasks and make a decent salary with it would be working for in average $2/hour on micro-task platforms. What my study did show though, is that at least to a certain degree the abilities of workers on MTurk and Upwork overlap and that some tasks that are offered on upper-end platforms can as well be performed by workers active on lower-end platforms (even without any supportive measures such as scaffolding). Considering the first research question which asks if scaffolding can enable novices to perform more complex task, strictly speaking it must be answered with an no. Novices did not need scaffolding to perform the complex tasks of this study at the same and in one case even higher level as Upwork workers. Scaffolding merely enabled them to do it significantly more satisfying for the requester.

Another surprising outcome of the study was that workers expressed higher satisfaction with their own work when working under the original condition. I do not have any evidence for a founded explanation of this phenomenon, but I have two hypotheses. First, I expect it to be easier to be satisfied with own performance when following an own approach and to satisfy personal expectations towards the outcome than to follow external guidelines and criteria. The criteria stated may not align with the personal taste, which makes it very difficult to be satisfied with the result even though they do meet the external expectations. Moreover, it can lead to a feeling of stress and dissatisfaction with the work itself as well when guidelines are followed because it must be checked constantly if no aspects have been forgotten or maybe even misunderstood or misinterpreted. My second hypothesis is based on the criteria I have chosen as the foundation for the evaluation of this thesis. As I already argued before, this list represents a personal taste. According to the criteria used in this study, workers performed significantly better when provided scaffolds. The list used for evaluating the submissions however, does not necessarily reflect the personal preferences and expectations of the workers. Their sets of criteria summarizing their subjective view may differ strongly from the one I have used, which means that according to their subjective view they may actually have performed better when solving the original task.
What is a little contrary to this result is the fact that workers rated the scaffolded task description to be slightly more satisfying than the original ones. When asked what in the task description workers found most helpful, which is the third research question, they named examples and step-by-step instructions. Here it seems as while workers appreciate a scaffolded task description, they may not be able to use it in a way that allows them to deliver results they are more satisfied with.

5.3. Implications

The results of this study indicate potential for possible changes on online labor platforms. First, this study indicates that requesters may underestimate the abilities of workers on micro labor platforms. Requesters invest resources into decomposition, allocation and re-integration of micro-tasks to reduce complexity and gain access to a wider pool of potential workers. However, in the study I conducted, MTurk workers were able to perform at least on par with professionals from Upwork working on complex tasks under the same conditions. As mentioned before, the tasks used in this study do not represent the upper-end of complexity on professional labor platforms, but they are tasks that occur frequently on the more professional platforms while they can almost not be found on lower-end labor markets. While these tasks may not be classic examples for tasks that would be decomposed and offered in smaller units on micro-task platforms, they represent a certain level of complexity that requesters are confident in trusting upper-end workers with while they rarely do so with lower-end workers. As this study shows, at least to some degree of complexity tasks that are currently almost exclusive to upper-end platforms can be handled by micro-task workers just the same as by professionals. Decomposing tasks into micro-units may therefore, at least in some cases, not be necessary and should be reevaluated for each case. Offering tasks on online labor platforms in a non-decomposed version would benefits both, workers and requesters. Workers could focus on one complex tasks instead of searching for multiple fitting micro-tasks, wasting their time on non-paid activity. Requesters on the other hand could avoid the investment of time and money for decomposing tasks into small units, searching for a fitting market for allocating the tasks and re-integrating the results. Additionally,
requesters using upper-end platforms may reconsider their choice of task allocation as MTurk workers are, at least for some work, equally qualified as workers on Upwork. Requesters may therefore use MTurk for getting their work done and thereby not just gain access to a new pool of potential workers but most likely save money by getting work done on a lower-end platform.

While micro-tasks and hence platforms for those would not completely disappear because not all micro-tasks are a part of a bigger, more complex problem, not all complex tasks can be solved without decomposition and not all requesters will stop decomposing tasks, there may still be a change in the market structure. An obvious scenario is that classic micro-task platforms extend their portfolio by more complex work. Since more complex work requires more time investment and is compensated with higher payment, the workers on MTurk would not need to constantly look for new work and could make more money. Considering the market structure, the balance of tasks on existing low and high-end platforms may remain as it currently is with the only change that some micro-tasks on low-end platforms are now offered as a complex version instead of multiple micro-tasks. However, since MTurk workers are used to lower payments they may accept lower wages for complex work than workers on upper-end platforms. Once the market realizes that complex work can be done successfully and for less money on lower-end platforms, requesters are likely to migrate their jobs to these markets or the prices on upper-end platforms will drop to the same level as on low-end platforms. It is quite probable that work that can be done on lower-end platforms will be offered there leaving upper-end platforms for work that is so complex, that it cannot be completed by micro-task workers. Furthermore, new crowd workers may be attracted due to interesting and decently compensated work expanding the online work market.

My study was based on research on the US online labor market. Crowd work, though, is especially interesting to international workers completing tasks for requesters from wealthier countries. While making $2/hour in the US is by far less than the minimum wage (United States Department of Labor 2017), more than a third of the world’s population lives on less than $2/day (Ruth 2012). Workers in the US may expect a higher payment for more complex work taking longer to complete, but foreign workers from
low-pay countries may be fully satisfied with a lower payment, which in their currency is still a reasonable salary. Scaffolding, which enhances work performance by explicitly stating expectations and guidelines how work should be done, may contribute to an increase of foreign and a decrease of domestic workers on online labor platforms in wealthier countries. A reasonable expectation would be that foreign workers provided with scaffolding task descriptions would successfully solve tasks on online labor platforms for wages that people from wealthier countries cannot compete with making paid crowd work a place that connects work from 1st world countries with workers from poorer environments.

Another possible scenario if the portfolio of lower-end platforms does not get extended by more complex work is that micro-task workers transition to higher-end platforms to gain access to “bigger” (timewise) and better pay tasks. Though, in my opinion this is a very unlikely scenario because in order to initiate such a transition, workers would have to realize that they are actually able to perform this kind of work. The results of my study indicate that to certain extent this is already the case and still, there is no significant movement from lower- to upper-end platforms.

While I have presented scaffolding as a method that has a generally positive impact on the performance on online labor markets there are still some thoughts that must be considered before committing to scaffolding. The first question requesters should ask themselves is if scaffolding is feasible for their particular task. While my research suggests that scaffolding increases the level of satisfaction with the outcome, it requires an investment of time and effort to design a scaffolded task. The effort to scaffold tasks is certainly one of the more important factors to consider. In my study I have chosen two tasks which I could solve based on my experience and general understanding of the topic. Though, I did not have any explicit knowledge about general criteria for the measurement of performance when solving such tasks. Therefore, I had to perform extensive research based on online media, requester reviews and academic literature gathering examples, summarizing best-practice, generating rubrics and learning to write in an engaging manner prior to scaffolding the tasks. As a result, and because I could not yet draw on
existing experience it took me in total 16 hours to finalize the first task description which was the description of the writing task.

Designing the second task, the slide-design task, took me only 7 hours, less than half the time. This significant decrease in time needed was at least partly based on the learning from scaffolding a task description once before, while it may as well have been partly due to the nature of the task or my pre-existing knowledge within this area. My experience was that scaffolding comes with a tradeoff over time. Advancing by practicing can be compared with learning how to ride a bike using training wheels. The more experienced requesters become, the more the time required diminishes. On the other hand, in practice it would be rather unacceptable to spend 7 hours on writing a task description for a slide design task. Here it must be considered that my high time investment was not just simply dedicated to writing and scaffolding the tasks. The expectations I had when designing the task description differ strongly from those of requesters. Since I was creating the foundation for a scientific study I had to consider every detail very carefully, guarantee to be as correct as possible and work accordingly to existing scientific standards. The task descriptions had to be designed in a way that lead to comparable submissions, which could be judged using the same criteria. Hence, rubrics, examples, step-by-step instructions had to be very specific to exclude innovative solutions that would not fit the criteria used and may hence be judge worse despite their good quality.

While requesters do not have to consider any of the aspects I just listed that were the main factors for my intense time investment in scaffolding, which may reduce the time they need for scaffolding significantly there are scenarios where scaffolding is still less advisable. These scenarios depend on the knowledge of the requester, the nature of task and the expectations towards the task. When the requester misses the understanding of the task and is unaware of what to expect from the solution scaffolding the task description will not have positive effects on the satisfaction with the outcome. The contrary may be the case as false rubrics may be communicated and workers may be misguided. Scaffolding even though in practice it may not require as much time as it did in this study it still goes along with some investment of time and effort. Hence, the extent (size and number of submissions needed) of the task plays an important role in the
decision if scaffolding should be applied. Scaffolding a task that can be completed in a shorter amount of time then required for scaffolding the description does not make sense. Accordingly, bigger tasks that the requester will pay more for and tasks which may be small but of which the requester needs a higher number of submissions are tasks predestined for scaffolding. The last factor which influences the decision if scaffolding should be applied is the expectation towards the task. My study indicated that by using scaffolding the satisfaction with performance can be increased by around 10%. This increase may not sound convincing enough to invest into scaffolding. When comparing the different submissions though, I realized that in my personal view basically only scaffolded submissions were acceptable. While this is a very subjective observation, it still points towards an important thought that should not be neglected. If these additional 10% in performance may enhance the performance from unacceptable to acceptable then scaffolding may be of even bigger value, then realized until now. Especially when dealing with tasks that are time critical or tasks for which feedback loops become easily expensive scaffolding might be a method worth investing in as it increases the chances of receiving acceptable outcome.

There are approaches which could make scaffolding more scalable and hence attractive to use. Crowd work platforms may get involved and support the development of a connected community that can exchange and profit from already existing tasks and scaffolded task descriptions. Another alternative is for platforms to instruct requesters how to write a scaffolded task description in order to achieve the best possible result. This could be realized via a layout that automatically forces requesters to use scaffolding as for example having a compulsory field where examples have to be posted or one with rubrics stating most important factors that should be considered when working and checked before the final result is submitted. Platforms could as well provide guidelines while requesters fill out the form for posting a task. Such a change is an additional one-time investment for platforms which, according to my study, would lead to significant improvement of work quality and hence customer satisfaction. An approach like this is already common practice in online environments where platforms support they customers, to achieve the best possible result. Examples are crowd funding platforms
(Kickstarter, GoFundMe, Indiegogo, etc.) where customers are advised how to present their project they are advertising best to raise most possible money or online application forums from universities (e.g. UC Berkeley) or incubator and accelerator programs (Y-Combinator, Startx, Amplify LA, etc.) where applicants are provided additional information how to fill out their application in order to be most informative and raise their chances for getting accepted.

5.4. Critical Evaluation of the Study

To provide a realistic view on the study and to classify the outcome and the suggestions appropriately it is particularly important to reflect upon the strength and weaknesses of this research. Hence strengths and weaknesses of the three aspects participants, tasks and data evaluation are presented.

Participants

To investigate the research questions stated in 1.3 this study has chosen workers from MTurk to be representative for low-end crowds and workers from Upwork as representatives for upper-end work. These two pools of workers were selected in order to express the dimension of distance between upper- and lower-end platforms and to investigate how close low-end workers could get to the performance offered on upper-end platforms. While both platforms are well known (MTurk internationally, Upwork at least in the US) and have a big pool of active workers, it is still questionable if investigating just two different platforms is representative for the whole spectrum of crowd workers. MTurk certainly stands for the low-end micro-task environment and the tasks offered there are typical for low-end platforms – tasks that require only basic cognitive skills. The variety of tasks offered on different upper-end labor markets though, is quite broad and can thereby not be fully expressed by just one platform as Upwork.

Considering the demographics of the participants I claim that the samples from MTurk and Upwork are fairly similar. The workers recruited on MTurk were 30% female and 70% male with an average age of 30.5 years (between 22 and 68 years old; over 80%
between 25 - 35 years old). Workers from Upwork were 33% female and 67% male with an average of 24.2 years (between 18 – 48 years old). The gender distribution is almost equal and taking into account the span of the working age on the US labor market that can be reached (18 years – 63, which is the average age of retirement (Backman 2016)) the 6.3 years of difference between the participant groups do not seem concerning. However, it still may have influenced the results. It may be that MTurk workers performed equal and even better than workers on Upwork because of over 6 more years of experience they have. What contradicts this hypothesis is that especially the slide-design task workers from MTurk performed significantly more satisfying. In my opinion however, younger people are more experienced in using programs such as PowerPoint. They are still closer to their school education where they have been thought to create presentations and where they practiced it in classes. What could provide additional insight and clarify some uncertainties are information about education and occupation. Unfortunately, these data were not raised in my study.

Furthermore, this study compared 114 MTurk workers with only 6 professionals. Obviously, the bigger the sample size, the higher the chances that the results are representative, and the conclusions derived from it have a certain validity. Hence, the small group of professionals whose work was evaluated in this study may not be representative for the average work on Upwork. However, it does at least provide a preview of what could be expected.

**Tasks**

This study was conducted with two exemplary tasks. The tasks were chosen based on how often they occur on Upwork, tasks from different fields, tasks with different required level of pre-existing knowledge and difficulty level that requires only entry level or intermediate skills to solve the task. This set of criteria was applied to guarantee the tasks are representative and appropriate for the goal of this study. Unfortunately, evaluating only two tasks may limit the informative value of the results. Tasks from other genres may be accessible to a different group of workers with different behavioral patterns that may influence their work performance. Moreover, the two tasks chosen do not represent
the whole spectrum of complexity. While micro-task workers were able to perform at least on par with upper-end workers when solving the tasks of this study, this may not have been the case if the tasks chosen were more complex tasks that require a more specified or broad set of skills. A further limitation is that the tasks I have used in this study are not task that would normally be decomposed into smaller units and offered as micro-tasks on low-end platforms, what is significant for the task I am targeting within this study. I argue though, that this is not a necessity for my study as my aim was to examine what level of complexity micro-task workers are able to handle (with and without scaffolding).

**Data Evaluation**

To guarantee an independent and objective evaluation of the data set, each submission was evaluated by two different judges specialized within the field of the specific task, blind to condition. To make the evaluation of the tasks comparable, a standardized list of criteria was applied. The list of criteria is based on previous academic efforts categorizing and summarizing most important aspects when solving these specific tasks as well as expectations extracted from best practice examples. However, the data evaluation does have some limitations. First and most important, the list of criteria, is not objective and not holistic. When speaking about evaluating tasks that have more than one correct outcome and in this case touch upon an artistic topic, the judgment criteria are always subjective as there is no such thing as a universal taste. Additionally, the judges evaluated the submissions based on the same list of criteria to minimize their personal influence and to standardize the judgement, still they did decide based on their very personal opinion if the submissions cover the criteria listed. Therefore, the evaluation cannot provide a general statement about quality and performance, but it can provide evidence of the degree to which the expectations of the requester, communicated via scaffolding and the evaluation criteria chosen, were met. Second, this study does not investigate the influence on performance of each (of the four techniques) used. This study can hence show that scaffolding did lead to more satisfying performance, but it cannot tell, which scaffolding methods contributed to which extent.
6. Conclusions and Future Outlook

Crowdsourcing provides employers with the possibility to outsource work on demand to a globally distributed workforce. It became a widely used source for human intelligence at reasonable cost which turned to a race of decomposition and cost saving. Platforms such as Amazon’s Mechanical Turk for example host mostly micro-tasks that pay no more than $0.10 each (Ipeirotis 2010b). Workers on this platform make an average salary of $2, which is fairly acceptable for workers from less wealthy countries but significantly below U.S. minimum wage (Ipeirotis 2010b; Agrawal et al. 2015; Ruth 2012). At the same time requesters criticize the low quality of work on micro-task platforms and struggle with decomposing complex task into micro-units, allocating the micro-tasks on suitable platforms and reintegrating the results to a complete solution. To offset this trend, this research considered how to help workers transition from micro-task work to being able to do rather complex contract job. It explored the use of proven techniques in learning science (examples, motivating task rationale, rubrics, step-by-step instructions) for scaffolding tasks within the micro-task paradigm. It investigated whether using these techniques can enhance performance and if so, to which extent.

It found that when applying the template of teaching techniques in a crowd context, micro-task workers perform significantly better, achieving results that met the expectations expressed by scaffolding. In a pairwise comparison experts chose scaffolded work from MTurk significantly more often than work from Upwork. Still, workers did not express any higher satisfaction with their results, even though they were of significantly higher quality. However, this may be due to the fact that workers judged their own work based on their personal taste which did not match the expectations and hence judgement criteria of the requester, leading to a different result. When given scaffolds, workers stated examples and step-by-step instructions to be most helpful for solving tasks.

Most surprisingly, the study shows that workers from MTurk, even without scaffolding, perform at least on par with workers from Upwork. In the writing task MTurk workers received a rating of 5.01 while Upwork workers received a rating of 5.00. In the slide-
design task workers from MTurk outperformed workers from Upwork significantly receiving a rating of 3.6 – over twice as high as the rating Upwork workers received (1.5).

The experiment provides evidence that the possibilities of crowd work may not be fully utilized in the current online labor market yet. As my study indicates, at least some work offered on high-pay online labor markets could as well be offered on lower-end platforms. If requesters desire outcome that meets their expectations better, scaffolding can be applied, however, it requires a certain knowledge of the task and a time investment and should hence be evaluated for each case whether it is feasible. To increase the general satisfaction with work performance platforms can invest in supporting their customers in scaffolding tasks. Possible approaches would be providing guidance, having a webpage layout that encourages scaffolding and automating knowledge and experience sharing on a platform for better and faster scaffolding at scale.

Future work could explore a wider range of tasks to represent a more accurate spectrum of work on online labor markets as well as different sets of criteria to validate the results of this study. Furthermore, the individual effect of the scaffolding methods on performance should be analyzed as well as the economic consequences going along with it. Methods to foster the development, testing, application, and sharing of scaffolding techniques on a big scale should be investigated since it would enable an organic development of the approach. Questions such as whether workers will apply for more complex tasks when these are scaffolded and if benefits from applying scaffolding prevail over the efforts associated with applying scaffolding require further analysis as well.
7. Declaration about the thesis

Ich versichere hiermit wahrheitsgemäß, die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, die wörtlich oder inhaltlich übernommenen Stellen als solche kenntlich gemacht und die Satzung des Karlsruher Instituts für Technologie (KIT) zur Sicherung guter wissenschaftlicher Praxis in der jeweils gültigen Fassung beachtet zu haben.

Karlsruhe, den 26.08. 2017

Vorname Nachname
Appendix

Appendix A

The writing task under scaffolded condition

To help other customers determine which mobile phone suits their needs best and may be the right choice for them we would like to gather product reviews of mobile phones. Therefore we would like you to write a review of a mobile phone of your choice (you may review your own phone). You are required to write a min. of 200 words following a “blogging style”. Listed below you find guidelines for how to complete this task successfully.

Before you start writing your review, you should prepare yourself and make sure that you:

Know the product: Be familiar with the product, it makes it easier to appear authentic and trustworthy.

Know the need it meets: Why would a customer by this particular product?

Know the customer: Who is the target group of the product (gender, interest, demographics, etc.) and which aspects are particularly interesting?

Writing the review follow this structure:

1. Introduction: Introduce the product as well as the promise it makes.

→ Engage the reader to continue reading

Examples:

“For a while now I’ve wanted to switch to an all natural shampoo, because the sulphates and other chemicals have been making my scalp itchy”

“I’ve been a fan of all of their other rose products — but this one really worked for me — and with a quickness. I started using this last Friday night”

2. Main body: Introduce the specific product, try to provide a descriptive visualization of the product, add practical details (price, size, weight, etc.), you may as well include a comparison to other products

→ Be a specific and detailed as possible

Examples:

descriptive visualization: “It smells good (although a little strong- not a kind to use in public) and is thick enough to soak deeply into my skin so I don’t need to reapply. I do find it a little too thick to use on my hands or when it’s just a normal day”

practical details: “This deodorant is all around solid. It’s unscented, isn’t loaded with chemicals and, at just over $3 per stick, is very affordable.”

comparison to other products: “If you have sensitive skin or prefer all-natural products, Soapwalla and Schmidt’s Natural Deodorant are standout performers”
3. **Verdict:** Provide a balanced evaluation of pro’s and con’s and an estimation either the product delivers its promise and summarize your statement

→ Be convincing without lobbying

Examples:

“I liked that it’s tinted, and self-tans. The self-tanning lasts a few days, as it’s very subtle, not the 2 weeks it claims. But the bronzing is immediate.”

“It’s a great choice if you’re looking for a natural men’s deodorant that not only works well but also smells great.”

Before you finalize your work revise your review checking if:

- it is specialized for the customer group considering all information they would be interested in
- it contains a balanced evaluation of all important aspects
- it comes to a clear conclusion
- it is catchy and if it sounds convincing and trustworthy
- the grammar and spelling is correct
Appendix B

The slide-design task under scaffolded condition
We would like your support in creating a 2-slide presentation of our company that shows: Who we are & What we do

Below you find a summary of the content we would like to be included as well as a link to our homepage providing you pictures of our team.

Link: https://moocita.org/
Content: We are a Non-Profit organisation. We help our students to acquire job relevant skills by providing online courses and assignments that are real world job tasks. Students solve these tasks alone or as teams and get paid if the quality of their work meets the necessary requirements. Working on real world tasks helps students to develop their portfolio and keep track of their progress. Instead of grades students see an actual impact on their bank account and receive letters of recommendation from employers.

When creating a powerful presentation, you should follow the guidelines listed below:

1. Understand the aim of the presentation and who the audience is
2. Once you start working on the presentation, make sure to include these rubrics:

Readability
- No spelling, grammar or punctuation mistakes
- Appropriate amount of text (no whole sentences)
- Font size is readable (even from distance)
- Line length and spacing are pleasant to read

Layout
- Fonts: fit the purpose, they can be read from distance, not too many different sizes
- Colors: support/highlight the content, but do not make it too colorful (distracting), have high contrast to the background
- Inserted Elements: high contrast between elements/text and background, margins around graphics, aligned to axis and other elements, use graphics to convey your message, not too many/too big inserted elements, not too much (if any) animation

Balance
- Symmetry is used to create a sense of order/movement/change

Simplicity
- Not too many information
- Bullet points instead of sentences
- No “complicated” background
- Enough white space (visual breathing room) on slide
Emphasis
Visual prominence to the most important information
Visual flow to guide viewers through the content
Visual proximity to related content
No overemphasis
Consistency
Clear organization or layout, information and design for viewers
Organizational scheme
Visual cohesiveness
s/unity
Elements that occupy similar positions in the information hierarchy should be
given similar graphic treatment
Appropriateness
The design successfully communicates a single clear message
The design successfully reaches out to the intended audience
The design successfully addresses the specified design brief

3. Before you submit your work, check again if you followed all rubrics

Example of a bad* slide:

Examples of good* slides:
Reference List

99DESIGNS. Available at: <https://99designs.com/> [Accessed 2 May 2017]


AMAZON MECHANICAL TURK. Available at: <www.mturk.com> [Accessed 04/25/17]


BRABHAM, D. C., 2010. *Crowdsourcing as a model for problem solving. Leveraging the collective intelligence of online communities for public good:* The University of Utah


CHAORDIX. Available at: <http://www.chaordix.com/> [Accessed 1 May 2017]


COOPER, S. et al., 2010. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307), 756–760

COURSERA. Available at: <https://www.coursera.org/> [Accessed 17 April 2017]


DOW, S. et al., eds., 2012. *Shepherding the crowd yields better work:* ACM


DUOLINGO. Available at: <https://www.duolingo.com/> [Accessed 21 April 2017]


EDWARDS, B. J. et al., eds., 2014. *Effects of Comparison and Explanation on Analogical Transfer:* Citeseer


FELSTINER, A., 2011. Working the crowd. Employment and labor law in the crowdsourcing industry

FISCHER, G. et al., eds., 1993. *Embedding computer-based critics in the contexts of design:* ACM

FREELANCER. *Hire expert freelancers for your job, online.* Available at: <https://www.freelancer.com/> [Accessed 1 May 2017]

GENTNER, D., LOEWENSTEIN, J., and THOMPSON, L., 2003. Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95*(2), 393


HAMMOND, J., and GIBBONS, P., 2005. What is scaffolding. *Teachers’ voices, 8*, 8–16


HERRINGTON, J., 2006. Authentic learning environments in higher education: IGI Global


INNOCENTIVE. Available at: <https://www.innocentive.com/> [Accessed 1 May 2017]


KHAN ACADEMY. Available at: <https://www.khanacademy.org/> [Accessed 27 April 2017]

KITTUR, A. et al., eds., 2013. The future of crowd work: ACM


LASECKI, W. S. et al., eds., 2015. *The effects of sequence and delay on crowd work: ACM*


LEE, B. et al., eds., 2010. *Designing with interactive example galleries: ACM*


LUTHER, K. et al., eds., 2015. *Structuring, aggregating, and evaluating crowdsourced design critique: ACM*


MARLOW, J., and DABBISH, L., eds., 2014. *From rookie to all-star: professional development in a graphic design social networking site: ACM*


NEBELING, M. et al., eds., 2016. WearWrite: Crowd-assisted writing from smartwatches: ACM


ONESPACE. Available at: <https://www.onespace.com> [Accessed 1 May 2017]


ROSS, J. et al., eds., 2010. Who are the crowdworkers? Shifting demographics in mechanical turk: ACM


SOLOLEARN. *Learn to code for free*. Available at: <https://www.sololearn.com/> [Accessed 27 April 2017]


individual case training. *Organizational Behavior and Human Decision Processes*, 82(1), 60–75

**TOYOTA MOTOR CORPORATION.** *History.* Available at: <http://www.toyota-global.com/showroom/emblem/history/> [Accessed 8 February 2017]


**UPWORK.** Available at: <www.upwork.com> [Accessed 25 April 2017]


**XU, A., and BAILEY, B.,** eds., 2012. *What do you think?: a case study of benefit, expectation, and interaction in a large online critique community:* ACM

**YUAN, A. et al.,** eds., 2016. *Almost an expert: The effects of rubrics and expertise on perceived value of crowdsourced design critiques:* ACM