Dipl.-Ing. Lisa Posch, BSc

Characteristics and Use of the International Microtask Workforce

Doctoral Thesis

to achieve the university degree of

Doctor of Technical Sciences (Dr. techn.)

submitted to

Graz University of Technology

Supervisor

Univ.-Prof. Dipl.-Ing. Dr. techn. Markus Strohmaier

Institute of Interactive Systems and Data Science

Graz, July 2020

Affidavit

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present doctoral thesis.

Date

Signature

Abstract

Advances in machine learning have automated many data analysis tasks. Nevertheless, there are still numerous problems that computers alone cannot yet solve and for which human input is still necessary. This need for human input in otherwise automated processes has created a new form of work: microtask crowdsourcing. On online platforms, researchers and businesses offer small tasks, called microtasks, to an anonymous, international crowd. People around the world then work on these microtasks, solving a multitude of digital problems that cannot yet be solved by automated methods alone.

This new form of work enables researchers and businesses to incorporate humans flexibly and seamlessly into otherwise automated systems. The platforms' interfaces abstract away all the human characteristics of the workers, and to the person requesting the work, receiving a worker's input may look identical to receiving the result of a method call. From the perspective of the worker, this type of work promises the ability to work from anywhere and the freedom to choose one's tasks and working hours. However, compared to traditional employment, it offers little social protection. There is no minimum wage, and workers are not entitled to any benefits such as vacation pay, sick leave, health insurance, or retirement benefits.

To better understand this emerging form of work, it is necessary to study the potentials of this new workforce in terms of *how* its input can be utilized to complement automated methods for data analysis. Furthermore, to gain a deeper understanding of this global phenomenon and its potential societal consequences, it is crucial to investigate *who* participates in the microtask workforce and *why* people around the world choose to participate in it. To that end, this thesis first demonstrates how human input from microtasks is complementary to automated methods in different stages of the machine learning process. Second, this thesis sets out to provide a comprehensive analysis of the characteristics of the international microtask workforce.

The first part of this thesis focuses on the use of the microtask workforce. It presents three use cases, in each of which the microtask workforce complemented automated methods in a different stage of the machine learning process. Focusing on the analysis of large text corpora, the use cases show how human input from microtasks complemented automated methods for the purposes of (i) evaluating a new topic model, (ii) analyzing populist political communication on social media, and (iii) evaluating different recommender algorithms with respect to their potential for incorporating information about users' current preferences.

The second part of the thesis focuses on the characteristics of the microtask workforce. It first provides an analysis of who participates in the international microtask workforce by presenting a comparative analysis of socio-demographic characteristics of workers in ten countries. Second, it presents the development and validation of the Multidimensional Crowdworker Motivation Scale, a theory-based and internationally applicable instrument for measuring motivations in the microtask context. Finally, to answer the question of why people around the world choose to participate in this type of work, this thesis presents a cross-country comparison of workers' motivations in ten countries.

The results presented in this thesis are relevant for researchers, practitioners, and policy makers interested in understanding this new form of work. Furthermore, the use cases presented in this thesis are relevant for data scientists concerned with the analysis of large text corpora.

Kurzfassung

In den letzten Jahrzehnten haben Fortschritte im Bereich des maschinellen Lernens viele Datenanalyseaufgaben automatisiert. Dennoch existieren weiterhin zahlreiche Probleme, die durch automatische Methoden allein noch nicht lösbar sind und für die nach wie vor menschlicher Input notwendig ist. Dieser Bedarf an menschlichem Input in ansonsten automatisierten Prozessen hat eine neue Form der Arbeit geschaffen: Microtask-Crowdsourcing. Auf Online-Plattformen veröffentlichen ForscherInnen und Unternehmen kleine Aufgaben, sogenannte Microtasks. Diese Microtasks werden dann von einem anonymen, internationalen Pool an Arbeitskräften bearbeitet.

Microtask-Crowdsourcing ermöglicht es ForscherInnen und Unternehmen, Menschen flexibel und nahtlos in ansonsten automatisierte Systeme einzubinden. Die Schnittstellen der Online-Plattformen verbergen alle menschlichen Charakteristika der ArbeiterInnen, und für den Auftraggeber macht es hinsichtlich des Ablaufs kaum einen Unterschied, ob das Ergebnis von einem Menschen oder von einem automatisierten System produziert wurde. Aus der Perspektive der ArbeiterInnen verspricht diese Art von Arbeit die Möglichkeit, von überall aus zu arbeiten und die Freiheit, Aufgaben und Arbeitszeiten frei zu wählen. Im Vergleich zu traditionellen Anstellungsverhältnissen bietet Microtask-Crowdsourcing jedoch wenig sozialen Schutz. Es existiert kein Mindestlohn, und die ArbeiterInnen haben keinen Anspruch auf Leistungen wie bezahlten Urlaub, Krankenversicherung oder Altersvorsorge.

Um diese neue Form der Arbeit besser zu verstehen, ist es zunächst notwendig zu untersuchen, wie menschlicher Input aus Microtasks automatische Methoden zur Datenanalyse ergänzen kann. Um ein tieferes Verständnis dieses globalen Phänomens und seiner potenziellen gesellschaftlichen Folgen zu erlangen, ist es darüber hinaus erforderlich zu untersuchen, wer in diese Art der Arbeit involviert ist und warum sich Menschen dafür entscheiden. Zu diesem Zweck beschäftigt sich diese Dissertation zunächst mit der Ergänzung automatischer Methoden durch menschlichen Input aus Microtasks und präsentiert dann eine umfassende Analyse der Charakteristika der Microtask-Arbeitskräfte.

Der erste Teil dieser Dissertation konzentriert sich auf den Einsatz der Microtask-Arbeitskräfte. Anhand von drei Anwendungsfällen wird gezeigt, wie menschlicher Input aus Microtasks automatische Methoden in verschiedenen Phasen des maschinellen Lernprozesses ergänzen kann. Die präsentierten Anwendungsfälle beschäftigen sich mit der Analyse großer Textkorpora und zeigen, wie menschlicher Input in verschiedenen Phasen des maschinellen Lernprozesses essenziell war für (i) die Evaluierung eines neuen Topic-Modells, (ii) die Analyse populistischer politischer Kommunikation und (iii) die Evaluierung von Recommender-Algorithmen hinsichtlich ihres Potenzials zur Einbeziehung von Informationen über aktuelle Benutzerpräferenzen.

Der zweite Teil der Dissertation beschäftigt sich mit den Charakteristika der Microtask-Arbeitskräfte. Zunächst wird ein Vergleich der soziodemografischen Charakteristika von Microtask-Arbeitskräften in zehn Ländern präsentiert. Darüber hinaus wird die Entwicklung und Validierung der Multidimensional Crowdworker Motivation Scale vorgestellt, einem theoriebasierten und international anwendbaren Instrument zur Messung von Motivationen im Microtask-Kontext. Um schließlich die Frage zu beantworten, warum sich Menschen für diese Art von Arbeit entscheiden, präsentiert diese Dissertation einen Vergleich der Motivationen von Microtask-ArbeiterInnen in zehn Ländern.

Die in dieser Dissertation vorgestellten Ergebnisse sind relevant für ForscherInnen, PraktikerInnen und politische EntscheidungsträgerInnen, die daran interessiert sind, diese neue Form der Arbeit zu verstehen. Darüber hinaus sind die in dieser Dissertation vorgestellten Anwendungsfälle relevant für DatenwissenschaftlerInnen, die sich mit der Analyse großer Textkorpora befassen.

Acknowledgements

First of all, I would like to thank my PhD supervisor, Markus Strohmaier, for his continuous support, invaluable advice, and inspiring discussions throughout the course of my doctoral studies. I am extremely grateful for his support, and it has been a privilege to work with him. Without his guidance and advice, this thesis would not have been possible.

I would like to express my profound gratitude to Arnim Bleier for providing frequent and thoughtful advice, feedback, and constant encouragement. My sincere thanks also go to Denis Helic for his much-valued scientific input and organizational support.

I am indebted to my co-authors for their valuable contributions and many fruitful discussions. Special thanks go to Sebastian Stier, Philipp Schaer, Clemens Lechner, Daniel Danner, Fabian Flöck, Lukas Eberhard, Simon Walk, Maryam Panahiazar, Olivier Gevaert, Michel Dumontier, Andreas Niekler, Christian Kahmann, Gregor Wiedemann, Kenan Erdogan, and Gerhard Heyer.

I would like to express my thanks to Claudia Wagner for her support and for creating a stimulating work and research environment, and I am indebted to Johann Schaible for all his support and patience. I would also like to extend my thanks to Mark Musen for the great discussions and for making me feel welcome during my research visit at the Stanford Center for Biomedical Informatics Research.

My thanks go to all colleagues and former colleagues in the Department of Computational Social Science at GESIS for the many exciting discussions and for their encouragement. In particular, I am grateful to Mohsen Jadidi, Lisette Munz, Anna Samoilenko, Indira Sen, Olga Zagovora, Nora Kirkizh, David Brodesser, Kathrin Weller, Haiko Lietz, Mattia Samory, Fariba Karimi, Juhi Kulshrestha, Marcos Oliveira, Mathieu Genois, Philipp Singer, Christoph Carl Kling, Florian Lemmerich, Roberto Ulloa, Maria Zens, Diana Lindner, Nadja Jelicic, Andreas Oskar Kempf, Agathe Gebert, Stefan Jakowatz, and Gerrit Hübbers. I would

also like to especially thank Jan Steinberg for being a great colleague and office partner. I would further like to express my gratitude to Rielies Neitzke for her valuable support in procuring books and to Karl Voit for providing the LaTeX template for this thesis.

I am indebted to the anonymous reviewers for their constructive feedback and insightful comments and suggestions. My sincere gratitude also goes to all the crowdworkers for their invaluable and essential contributions to the work presented in this thesis.

I owe my deepest gratitude to my parents, Hermine and Karl, who provided unconditional support and encouragement during my studies. Finally, I am very grateful for all the encouragement, moral support, and feedback I received from my friends.

Thank you!

Contents

1 Introduction

1.1 Motivation for this Thesis

Automation is replacing certain forms of human labor. At the same time, however, it is creating a large demand for new types of digital human work. In the past decade, industry and academia alike have increasingly made use of a new type of workforce for those types of digital labor that cannot be performed by computers alone. In this new type of work, small, self-contained tasks called *microtasks* are outsourced to a large crowd of workers, often from geographically, economically, and culturally diverse backgrounds. This crowd, an "indefinite and unknown" (see, e.g., Mandl et al., 2015) pool of human workers, is accessible via online platforms where workers can register to perform these tasks in exchange for payment.

Many processes that seem automated to the onlooker rely, in reality, on this large, indispensable human workforce behind the scenes. Tasks such as filtering undesired content like hate speech on social media, tagging objects in images, collecting and verifying data from the web, or removing near-duplicate listings in a database often still rely on human labor. In many cases, machine learning methods still require human input in different stages of the machine learning process. For example, training and test datasets have to be created by humans before supervised machine learning models can be trained and evaluated, and unsupervised models often need to be evaluated via a process that involves human input.

This necessity for human input in different stages of the machine learning process is unlikely to disappear in the near future. Even though advances in machine learning continue to automate many data analysis tasks, new solutions often give rise to new opportunities for automating another task. We therefore continually identify new tasks that currently require human labor but have the potential to be automated. Gray and Suri (2019) call this phenomenon the *"paradox of automation's last mile."* As microtask platforms provide the flexible workforce that is needed to perform many of the tasks that cannot yet be performed by computers alone, this type of work is likely here to stay.

Work on microtask platforms fits in with a wider trend towards increasingly flexible and shorter-term work arrangements that has been observed in industrial societies (see, e.g., Kalleberg, 2009; Hewison and Kalleberg, 2013). The precarious nature of work on microtask platforms has led to policy discussions around working conditions and social protection of workers (see, e.g., European Parliament, 2016; Waas et al., 2017). To inform such discussions, it is crucial to gain a better understanding of this emerging form of work.

This thesis sets out to deepen our understanding of work on microtask platforms and of the international workforce involved in it. First, this thesis demonstrates *how* human input from the crowd is complementary to automated methods in different stages of the machine learning process. Second, this thesis provides analyses of *who* participates in the international microtask workforce and *why* people around the world choose to participate in it.

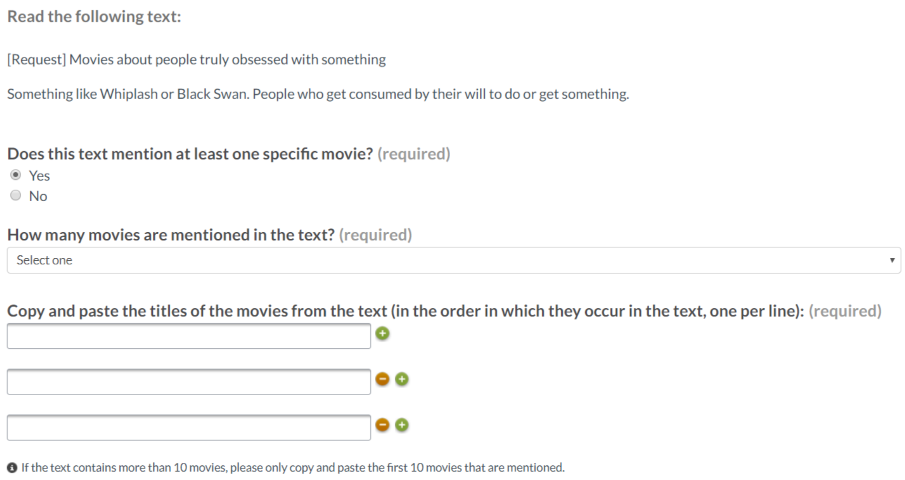
The remainder of this chapter first gives an introduction to the concepts of microtasks and microtask platforms in Section 1.2. Section 1.2 further provides an introduction to the general characteristics and implications of this new form of work. Then, in Section 1.3, this chapter presents the overall problem statement, objectives, and the general approach of this thesis. Section 1.4 presents the research questions addressed in this thesis, including an overview of the respective problems, approaches, and findings. Section 1.5 provides a list of the publications contained in this cumulative thesis, and Section 1.6 gives an overview of the main contributions and implications of this work. Finally, Section 1.7 gives an overview of the general structure of this thesis.

1.2 Microtasks

The term *crowdsourcing*, a portmanteau of the words "crowd" and "outsourcing," was introduced by Jeff Howe, who defined crowdsourcing as *"the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call"* (Howe, 2006). Similarly, but explicitly including only those activities that are performed in exchange for payment, a Eurofound report by Mandl et al. (2015) defined the term *crowd employment* as a type of employment that *"uses an online platform to enable organisations or individuals to access an indefinite and unknown group of other organisations or individuals to solve specific problems or to provide specific services or products in exchange for payment."*

Microtask crowdsourcing is a type of crowdsourcing where very small tasks, called *microtasks*, are offered online to an anonymous crowd to be worked on. Usually, workers are paid on a per-task basis, and a single task typically pays only a few cents upon completion. Completing a microtask generally takes only a few minutes, and in many cases, only seconds. Organizations or individuals who request the work are usually called *requesters*, and the workers who complete the tasks are called *crowdworkers* or *microworkers*.

Typical microtasks are tasks that are generally easy to do for humans but hard to do for computers, in the sense that no efficient or accurate algorithm has been developed for the task so far. Most microtasks rely on general human cognitive abilities and do not require any specialized knowledge from the crowdworkers. For example, microtasks may be employed for identifying the sentiment expressed in short snippets of natural language text, for the categorization of product images, for



Read the following text:

[Request] Movies about people truly obsessed with something

Something like Whiplash or Black Swan. People who get consumed by their will to do or get something.

Does this text mention at least one specific movie? (required)

● Yes
○ No

How many movies are mentioned in the text? (required)

Select one ▾

Copy and paste the titles of the movies from the text (in the order in which they occur in the text, one per line): (required)

If the text contains more than 10 movies, please only copy and paste the first 10 movies that are mentioned.

Figure 1.1: **Example of a Microtask.** This figure shows an exemplary microtask in which workers are asked to identify movie titles in a text.[1]

identifying adult content, or for data matching. Figure 1.1 shows an exemplary microtask in which workers are asked to identify movie titles in a given unstructured text. In many cases, microtasks are incorporated into otherwise automated workflows, and requesters often offer large batches of identically structured microtasks to the crowd. A worker may repeat the same type of task hundreds or even thousands of times.

Jeff Bezos, CEO of Amazon, described the advantage of using microtasks in the following way: *"Just as you would write any subroutine in code, you can now write a subroutine that will go out and get an answer for you [from the human crowd]"* (Bezos, 2006). Alternative terms for the concept of microtasks, such as "human intelligence tasks" (HITs) or "artificial artificial intelligence," (see, e.g., Bezos, 2006; Amazon Mechanical Turk, 2016) also highlight that human intelligence is being used as a substitute for software algorithms, for those tasks that artificial intelligence cannot yet solve satisfactorily. In an interview with The New York Times, Bezos explained: *"Normally, a human makes a request of a computer, and the*

[1] The microtask shown in Figure 1.1 was used in Eberhard et al. (2019), and the text shown in the example is taken from `https://www.reddit.com/r/MovieSuggestions/comments/3bf1yo`.

computer does the computation of the task. But artificial artificial intelligences like Mechanical Turk invert all that. The computer has a task that is easy for a human but extraordinarily hard for the computer. So instead of calling a computer service to perform the function, it calls a human" (Pontin, 2007).

1.2.1 Microtask Platforms

Microtask platforms act as intermediaries between task requesters and crowdworkers. The platforms provide task requesters with the necessary infrastructure for designing and publishing microtasks, and they provide workers with the infrastructure to select tasks, to work on their chosen tasks, and to submit their work. Platforms also generally handle at least part of the payment process, subtracting a commission for their service.

After registering an account on a platform, workers are offered a list of tasks that are available for them to work on. The platform displays the task to the worker, who can then work on it and submit the finished work to the platform. The task requester then receives the results via the platform's web interface or API. Requesters and workers typically interact with each other exclusively via the platform. In most cases, there is only a one-way communication from the requester to the worker via written task instructions.

Publicly launched in 2005, Amazon Mechanical Turk (MTurk)[2] was the first microtask platform (Amazon Mechanical Turk, 2015). The platform was named after the Mechanical Turk, a chess-playing machine constructed in the eighteenth century. This chess-playing machine was presented to the audience as an automaton that was able to play chess of its own accord (Hindenburg, 1784), but in reality, it contained a hidden human chess player who controlled the machine's moves. In 2007, the microtask platform Figure Eight[3] was founded by Lukas Biewald and Christopher Van Pelt as "Dolores Labs" (Barret, 2009). The platform soon

[2] https://www.mturk.com/

[3] https://www.figure-eight.com/

changed its name to "CrowdFlower" (Rao, 2009; Ha, 2012), and in 2018, it was renamed again to "Figure Eight" (Figure Eight, 2018b). In 2019, Figure Eight was acquired by the company Appen[4] (Appen, 2019).

Together, MTurk and Figure Eight have been estimated to share about 80% of the microtask market, with approximately equal revenues (Kuek et al., 2015). Besides these two large platforms, there are also numerous smaller microtask platforms, such as Clickworker[5], Microworkers[6], and Crowdee[7].

While the general functionality and workflow is similar across different microtask platforms, the specific implementations and features provided may differ. For example, the platform MTurk does not offer any built-in functionality for assessing workers via test questions, whereas the platform Figure Eight encourages requesters to upload a gold standard against which workers are then continually assessed during their work. Other differences in functionality include, for example, different graphical interfaces or markup languages for implementing tasks, different ways of handling how workers are paid, different ways of recruiting workers, and different reputation systems for workers. A detailed list of features that may differ across platforms can be found in Vakharia and Lease (2015).

Furthermore, platforms may differ in the workforce that they attract. For example, the platform Figure Eight attracts a much more international workforce than MTurk. The vast majority of MTurk's workforce consists of workers located in the United States and India, most likely due to the restrictive payment options that the platform offers in other countries.[8] By contrast, parts of the payment process on the platform Figure Eight

[4] `https://appen.com/`

[5] `https://www.clickworker.de/`

[6] `https://www.microworkers.com/`

[7] `https://www.crowdee.com/`

[8] While workers in the United States and in India can receive local currency for their work, amazon.com gift cards are the only payment option for workers located in other countries (Amazon Mechanical Turk, 2018). However, the demographics of the platform's workforce might change in future due to MTurk recently enabling payments

are handled by independent partner websites, which provides workers with much more flexibility regarding the country and currency of their payment.

1.2.2 A New Form of Work

Microtask platforms provide businesses and other microtask requesters with an unprecedented opportunity to access a global workforce on demand. Being an extremely short-term and flexible type of work, work on microtask platforms fits in with a wider trend that has been observed in industrial societies, a trend towards increasingly flexible work arrangements that are characterized by short-term, market-based contracts (see, e.g., Kalleberg, 2009; Hewison and Kalleberg, 2013).

As Lukas Biewald, co-founder of the microtask platform Figure Eight, stated in 2010: *"Before the Internet, it would be really difficult to find someone, sit them down for ten minutes and get them to work for you, and then fire them after those ten minutes. But with technology, you can actually find them, pay them the tiny amount of money, and then get rid of them when you don't need them anymore"* (quoted in Marvit, 2014, also see, e.g., De Stefano, 2016).

This new form of work enables microtask requesters to seamlessly incorporate humans into software procedures, with the "indefinite and unknown" crowd of workers being accessible on demand via the interfaces of microtask platforms. The platforms' interfaces abstract away all the human characteristics of the workers, and the workers' recruitment, payment, and the evaluation of their work are automatically handled by software. To the person requesting the crowd's labor, receiving input from a human may look no different than receiving the result of a method call. Jeff Bezos, CEO of Amazon, described the intended purpose of the microtask platform Amazon Mechanical Turk in the following way: *"You've heard of software-as-a-service. Well this is basically people-as-a-service."*

in US$ for workers in 25 countries outside the U.S., provided that they have a U.S. bank account (Amazon Mechanical Turk, 2019).

(Bezos, 2006, also see Irani, 2015b). The view that humans represent an "artificial" form of software algorithms in the systems that use their labor is also reflected in the term "artificial artificial intelligence" that the platform uses for describing this type of work.

Such an automaton-like view of humans has been termed "mechanistic dehumanization" (Haslam, 2006). When humans are denied their human attributes and perceived as being equivalent to machines, they are perceived to be fungible, i.e., interchangeable with others of their type, and lacking in individual agency, emotionality, and other attributes that define human nature (Haslam, 2006). Consequently, they may be treated with indifference and disregard, and as a means to an end (Haslam, 2006; Bastian and Haslam, 2011).

A number of authors have raised concerns regarding the dehumanizing nature of work on microtask platforms (see, e.g., Ross et al., 2010; Kittur et al., 2013; Irani, 2015a; De Stefano, 2016; Berry, 2019; Gray and Suri, 2019; Barbosa and M. Chen, 2019). For example, Ross et al. (2010) noted that *"obscuring worker identity may [...] potentially contribute to workers being exploited: because workers are decontextualized, requesters may be more likely to offer lower, unfair prices on HITs, or even refuse to pay for work performed."* De Stefano (2016) argued that the concept of humans-as-a-service *"perfectly conveys the idea of an extreme form of commodification of human beings."* Emphasizing the invisibility of crowdworkers to those who benefit from their work, Gray and Suri (2019) termed work on microtask platforms "ghost work" (also see Marvit, 2014). This invisibility, Gray and Suri (2019) argued, can *"make requesters forget they are even hiring humans."*

While the technology-enabled aspects and the global scale of work on microtask platforms are genuinely new, some authors have pointed out that this is not true of all aspects of this type of work (see, e.g., Felstiner, 2011; De Stefano, 2016; Cherry, 2016; Finkin, 2016; Waas et al., 2017). In many ways, work on microtask platforms resembles industrial homework, where workers produce goods in their homes and are paid on a piecework basis (Finkin, 2016; Waas et al., 2017). Due to this resemblance, the terms "cognitive piecework" and "digital piecework" have sometimes been used to describe work on microtask platforms (see, e.g., Felstiner, 2011; Fieseler et al., 2019). Cherry (2016) also pointed out that *"breaking down tasks to their lowest common denominator"* is not a new idea and argued that this aspect of work on microtask platforms resembles the *"de-skilled industrial processes associated with Taylor, but without the loyalty and job security."*

Compared to traditional employment, work on microtask platforms is subject to little legal regulation. There are no work contracts involved, besides the workers' choice to accept or reject a microtask platform's terms of service. Microtask platforms generally view their workers as independent contractors (see, e.g., Cherry, 2016; Waas et al., 2017) and emphasize in their terms of services that workers are not considered employees (see, e.g., Figure Eight, 2018a; Amazon Mechanical Turk, 2020). Workers are therefore not entitled to any benefits that a person considered an employee would be entitled to, such as vacation pay, sick leave, health insurance, or retirement benefits.

Minimum wage laws do not apply, and estimates of the hourly wage that workers achieve on micro-task platforms range from under US$1 to around US$5 (see, e.g., T. Kaplan et al., 2018; Berg, 2015; Ross et al., 2010; Horton and Chilton, 2010; Khanna et al., 2010). From this income, workers have to supply their own tools and office space, and they have to pay for the internet access needed to perform the work. Workers further bear the risk of any software implementation errors in the microtasks or the platform, which can result in workers not being paid for completed work. Moreover, a worker's account may be suspended by the platform at any time, with the worker having little to no recourse to appeal the decision (see, e.g., Gray and Suri, 2019).

On the other hand, the flexibility of this type of work may also provide advantages for workers. Kessler (2014) phrased the vision of crowd employment in the following way: *"Whatever you do, it will be your choice. Because you are no longer just an employee with set hours and wages working to make someone else rich. In the future, you will be your very own mini-business."* The promise to workers is that they will have the autonomy to decide where and when they want to work, and which tasks they want to accept. The ability to work from anywhere may also offer an opportunity to generate income for people who would struggle to work outside their home, for example due to disabilities (see, e.g., Berg, 2015). When describing the benefits of work on microtask platforms, many workers state that they appreciate this flexibility (see, e.g., Deng and Joshi, 2016; Berg et al., 2018).

However, De Stefano (2016) and others (see, e.g., Pesole et al., 2018; Prassl, 2018; Gray and Suri, 2019) have argued that these beneficial aspects should not be over-estimated. First of all, the global competition between workers leads to downward pressure on wages, which reduces the workers' flexibility as they may have to work for many hours to generate any significant income (De Stefano, 2016). Furthermore, while workers can theoretically choose their working hours, the reality is often different. In practice, workers who rely on this income are often forced to continuously monitor the stream of new tasks being posted and spend many unpaid hours looking for suitable tasks. Otherwise, they risk losing important income (see, e.g., Gray and Suri, 2019). Moreover, the best-paying tasks may be posted by companies in a different time zone, which further limits the workers' flexibility and may force workers to stay alert and work on tasks during the night (see, e.g., Gray and Suri, 2019; Gupta, Crabtree, et al., 2014).

Despite the seemingly large disadvantages compared to traditional employment, especially for workers in high-income countries, a large global microtask workforce has emerged during the past decade. The growth of crowd employment, including work on microtask platforms, has given rise to policy discussions on social protection and working conditions of crowdworkers (see, e.g., Felstiner, 2011; European Parliament, 2016; European Commission, 2016; Codagnone et al., 2016; Waas et al., 2017). The discourse around this type of work includes discussions on whether crowdworkers should be considered employees rather than independent contractors, and whether a new category of employment might be needed to adequately regulate crowd employment. There is also an ongoing discussion on whether this type of work should be considered "work" at all or whether it should be considered a spare-time activity, with remuneration playing only a minor role for workers (see, e.g., European Parliament, 2016; Berg, 2015).

To achieve a better understanding of this emerging form of work, it is important to not only understand for what purposes the workforce is being used, but also to understand who, around the world, participates in it and why. To that end, this thesis sets out to provide a detailed picture of the characteristics and different uses of the international microtask workforce.

1.3 Problem Statement, Objectives, and General Approach

Problem Statement. In the past decades, advances in machine learning have automated many data analysis tasks. However, current methods have limitations that can only be overcome by incorporating human labor into the otherwise automated processes. This need for flexible, on-demand human labor, along with increasing worldwide internet access, has brought forth a new type of global workforce. On microtask platforms, workers from around the world work on solving a multitude of digital problems that cannot yet be solved by automated methods alone.

To understand this emerging form of work, it is necessary to study the potentials of this new workforce in terms of *how* its input can be utilized to complement automated methods for data analysis. Furthermore, to gain a deeper understanding of this global phenomenon and its potential societal consequences, it is crucial to investigate *who* participates in the microtask workforce and *why* people around the world choose to participate in it. So far, most research regarding who participates in the microtask workforce has focused on workers from only two countries, the United States and India. However, work on microtask platforms is a global phenomenon, and little is known about the workforce in other countries. Furthermore, there is little knowledge of the motivations that people around the world have for participating in this new form of work, and there is currently no theoretically founded and well-validated instrument for comprehensively measuring motivations in the microtask context.

Objectives. The overarching objective of this thesis is twofold: First, this thesis sets out to demonstrate ways in which human input from the crowd is complementary to automated methods for data analysis, in different stages of the machine learning process. Specifically, this thesis focuses on employing microtasks for complementing automated methods for the analysis of large text corpora. Second, this thesis aims to provide a detailed picture of the characteristics of the international microtask workforce, in an attempt to shed light on the human attributes that have been abstracted away from this global, "indefinite and unknown" pool of human workers. To that end, this thesis sets out to conduct the first comprehensive, large-scale comparative analysis of socio-demographic characteristics of crowdworkers in different countries that goes beyond an analysis of workers located in the USA and India. Additionally, this thesis sets out to develop a theoretically founded and internationally applicable instrument for measuring the motivations of the microtask workforce, with the aim of conducting the first cross-country comparison of crowdworkers' motivations to participate in this type of work.

General Approach. The first part of this thesis employs microtasks to complement automated methods in different stages of the machine learning process, addressing the question of *how* human input from microtasks can complement methods for the analysis of large text corpora. The second part of this thesis employs microtasks to analyze the socio-demographic characteristics and motivations of the international microtask workforce, addressing the questions of *who* chooses to participate in this type of work and *why* people around the world choose to participate in it. To measure the different concepts of interest, this thesis makes use of a range of methods for estimating latent variables in structured and unstructured data.

1.4 Research Questions

This thesis sets out to provide a better understanding of the international microtask workforce by analyzing *how* the microtask workforce can be used to complement automated methods, *who* chooses to participate in it, and *why* people choose to participate in it.

The first part of this thesis focuses on the use of the microtask workforce, specifically regarding *how* microtasks can be employed in different stages of the machine learning process to complement automated methods of text analysis. Thus, the overarching research question for the first part of this thesis is the following:

RQ1: How can human input from microtasks complement methods for the analysis of large text corpora in different stages of the machine learning process?

The second part of this thesis focuses on answering questions related to understanding the characteristics of the international microtask workforce, specifically regarding *who* participates in the workforce and *why* people participate in it. Thus, the overarching research questions for the second part of this thesis are the following:

RQ2: What are the socio-demographic characteristics of the international microtask workforce, and do these characteristics differ across countries?

RQ3: Why do people choose to participate in the microtask workforce, and do their motivations differ across countries?

This section introduces these research questions in detail and describes the approach used to address each question as well as the main findings and contributions.

RQ1: How can human input from microtasks complement methods for the analysis of large text corpora in different stages of the machine learning process?

Problem. This research question aims at understanding how the microtask workforce can be used in machine learning problems, with a focus on the analysis of large text corpora. Methods for the automated analysis of text have limitations in different stages of the machine learning process, and this thesis demonstrates how different limitations can be overcome by incorporating human input from microtasks. The specific problems we aimed to address by complementing automated methods with human input from microtasks were the following:

- How does the semantic coherence of the newly developed Polylingual Labeled Topic Model compare to that of existing topic models? (Posch, Bleier, Schaer, et al., 2015)
- How do political actors in Germany differ with respect to their use of populist communication? (Stier, Posch, et al., 2017)
- To what extent can information contained in narrative descriptions of users' current preferences help to improve the recommendations of established recommender algorithms? (Eberhard et al., 2019)

Approach. To demonstrate how necessary human input can be provided via microtasks, this thesis presents three different use cases, in each of which the microtask workforce was involved in a different stage of the machine learning process. A high-level view of the machine learning process is depicted in Figure 1.2 (Section 1.7). Specifically, in Article 1 (Posch, Bleier, Schaer, et al., 2015), presented in Section 3.2.1, we employed the microtask workforce in the *model evaluation* stage, to evaluate the semantic coherence of a new topic model by comparing it to the semantic coherence of three existing topic models. In Article 2 (Stier, Posch, et al., 2017), presented in Section 3.2.2, the microtask workforce was involved in the *model interpretation* stage of the machine learning process, by interpreting a topic model's parameters with respect to populist communication. In Article 3 (Eberhard et al., 2019), presented in Section 3.2.3, the microtask workforce contributed to the *data preparation and preprocessing* stage of the machine learning process, performing tasks such as sentiment analysis and the extraction of important information from unstructured text. Additionally, in Stier, Bleier, Bonart, Mörsheim, et al. (2018b)[9], the microtask workforce was involved in the *data collection* stage, by collecting social media accounts of mainstream as well as alternative German media on Facebook and Twitter.

Findings and contributions. In each of the individual research projects, a different limitation of an automated method was overcome by introducing human input from microtasks. In each case, this human input was indispensable for addressing the respective problem and answering the project-specific research questions. In the following, I give an overview of the research projects that employed the microtask workforce to complement automated methods. For each project, I describe in which stage of the machine learning process microtasks were used, which concrete tasks crowdworkers performed to complement automated methods, and what the main contribution of the project was.

[9] While I was responsible for the design, implementation, and execution of all microtasks in this project, my contribution to the project constituted a comparatively small part in a large research collaboration. This publication is therefore not included in this cumulative thesis. The result of this research effort is a dataset (Stier, Bleier, Bonart, Mörsheim, et al., 2018a) that enables researchers to study online political communication in Germany.

- **Article 1: The Polylingual Labeled Topic Model** (Posch, Bleier, Schaer, et al., 2015)

  Stage: Model evaluation

  Microtasks: Evaluate the semantic coherence of topics estimated by the newly developed PLL-TM, compared to existing models.

  Contribution: The PLL-TM, a new topic model for estimating topics in multilingual, labeled documents. A visualization system based on the PLL-TM was published separately (Posch, Schaer, et al., 2016).

- **Article 2: When Populists Become Popular: Comparing Facebook Use by the Right-Wing Movement Pegida and German Political Parties** (Stier, Posch, et al., 2017)

  Stage: Model interpretation

  Microtasks: Interpret model parameters in the context of populist communication.

  Contribution: An analysis of populist political communication on social media by German political actors.

- **Article 3: Evaluating Narrative-Driven Movie Recommendations on Reddit** (Eberhard et al., 2019)

  Stage: Data preparation & preprocessing

  Microtasks: Extract relevant information from unstructured text, sentiment analysis.

  Contribution: An evaluation of recommender algorithms with respect to their potential for incorporating information contained in narrative descriptions of users' current preferences.

- **Systematically Monitoring Social Media: The Case of the German Federal Election 2017** (Stier, Bleier, Bonart, Mörsheim, et al., 2018b)

  Stage: Data collection
  Microtasks: Collect social media accounts (from Facebook and Twitter) of mainstream and alternative media.
  Contribution: A dataset (Stier, Bleier, Bonart, Mörsheim, et al., 2018a) that enables researchers to study online political communication in Germany.

RQ2: What are the socio-demographic characteristics of the international microtask workforce, and do these characteristics differ across countries?

Problem. Research on the socio-demographic characteristics of the microtask workforce has almost exclusively focused on the two countries that constitute MTurk's target audience, i.e., the USA and India. So far, little is known about the microtask workforce on other platforms and in countries other than the USA and India. However, work on microtask platforms is a global phenomenon, and the platform Figure Eight, the second market leader in the microtask market (Kuek et al., 2015), targets a much more international audience than MTurk (see, e.g., Berg, 2015). This research question aims at complementing existing literature by providing a more comprehensive picture of the international microtask workforce regarding socio-demographic characteristics of workers in different countries and regarding the importance that the income from microtasks has in the workers' lives.

Approach. To gain insights into the socio-demographic characteristics of the international microtask workforce, in Section 3.3.1, this thesis presents a large survey of crowdworkers in ten different countries and at two points in time (Article 4, Posch, Bleier, Flöck, et al., 2018). The survey was conducted on the platform Figure Eight, and we collected data from 900 workers in each country at each time point, for a total of 18,000 responses. We selected the countries from diverse income levels and additionally aimed for a high cultural diversity as well as sufficient activity on the platform. Furthermore, to capture a diverse sample of workers in each country, we split the starting times of the tasks into three groups: typical working hours and evenings in the respective time zones, and weekends. The survey included questions regarding different socio-demographic characteristics of the workers as well as questions regarding the importance of microtask income for the workers' lives. This approach allows us to not only compare the characteristics of different countries' microtask workforces, but also to analyze their stability over time by calculating the Jensen–Shannon divergences (Lin, 1991) between two independent samples taken eight months apart.

Findings and contributions. The results of this analysis provide a detailed picture of the international microtask workforce in ten countries. The analysis constitutes the first large-scale country-level comparison of socio-demographic characteristics of the microtask workforce that goes beyond an analysis of U.S.-based and Indian workers on the platform MTurk. The results of the analysis revealed wide-ranging differences regarding the demographic composition, time spent on the platform, reliance on microtask income, and use of microtask income between the different countries. Furthermore, the results showed that these characteristics remained largely stable between the two independent samples collected at different points in time.

RQ3: Why do people choose to participate in the microtask workforce, and do their motivations differ across countries?

Problem. The question of why people around the world choose to participate in the microtask workforce still remains largely open. While there has been some research on the motivations of the microtask workforce, it has, like research on the socio-demographic characteristics, focused mainly on the two countries that constitute MTurk's target audience. Most importantly, however, even for workers on MTurk, there is currently no well-validated, theoretically founded instrument for measuring different types of motivations in the microtask context. Moreover, measuring the motivations of the microtask workforce in different countries requires the measurement instrument to be valid in each country, and any cross-group comparisons of motivations (such as comparisons between countries) additionally require the measurement instrument to be invariant across the groups of interest. Thus, answering the overarching research question *RQ3* requires first answering the following research questions:

RQ3.1: How can we validly measure motivations in the microtask context?

RQ3.2: Is the instrument used to measure motivations in the microtask context applicable in different countries?

RQ3.3: Is the instrument used to measure the motivations of the microtask workforce suitable for conducting cross-country comparisons?

Approach. Article 5 (Posch, Bleier, Lechner, et al., 2019), presented in Section 3.3.2, tackles these research questions within the framework of self-determination theory (SDT), a theory of human motivation that has been successfully applied to measure work motivation in the traditional employment context (see Section 2.2).

To address *RQ3.1*, we first conducted an evaluation of the suitability of two established SDT-based work motivation scales that were developed for the traditional employment context. We performed minimal adaptations to the item wordings in order to semantically adapt them to the microtask context, and then conducted confirmatory factor analyses (see Section 2.1.2) to evaluate different measurement models based on data collected from workers in the USA.

Based on the results of these analyses, we conducted exploratory factor analyses to identify which parts of the traditional work motivation scales were potentially useful for measuring motivations in the microtask context. To develop an instrument for measuring motivations in the microtask context, we then compiled an item pool, which we reduced and refined by conducting exploratory factor analyses on data collected from workers in three culturally diverse countries.

With the final 18-item version of the *Multidimensional Crowdworker Motivation Scale* (MCMS), we collected data from ten countries, which we selected for cultural diversity and from different World Bank income groups. We evaluated the internal consistency of the different motivational dimensions and conducted confirmatory factor analyses to evaluate the construct validity of our hypothesized six-factor model. Given the good model fit overall as well as adequate fit in all income groups and countries, we further evaluated additional aspects of the model's validity. The results of these analyses provided evidence that the MCMS constitutes a reliable and valid measurement of motivations in the microtask context, thus answering *RQ3.1* and *RQ3.2*. To answer *RQ3.3*, we conducted measurement invariance tests, the results of which indicated that partial scalar invariance holds between countries and between income groups, allowing for valid cross-group comparisons of latent means. Having answered *RQ3.1*, *RQ3.2*, and *RQ3.3*, we could then proceed to answer the overarching *RQ3* by comparing the model-estimated latent means of workers in different counties.

Findings and contributions. The main contributions of the work presented in Section 3.3.2 are twofold. First, it presents the *Multidimensional Crowdworker Motivation Scale*, a theoretically well-founded, validated, and internationally applicable instrument for measuring motivations in the microtask context. Second, it presents an analysis and comparison of workers' motivations in ten different countries and three country income groups.

The results of the comparison showed both similarities and significant differences between the countries and income groups. For example, material external regulation was the motivational dimension with the highest mean overall as well as in all countries and income groups, closely followed by intrinsic motivation. This indicates that both monetary rewards and enjoyment inherent in the activity play an important role for crowdworkers around the world. However, both monetary rewards and enjoyment were somewhat more important to workers in middle- and low-income countries than to workers in the USA and in Germany, whereas workers in the USA and in Germany exhibited, on average, a higher lack of motivation. Furthermore, the results indicated that, in all groups, putting effort into microtasks was moderately in alignment with workers' personal goals such as lifestyle preferences or career goals, but that this dimension was less important in high-income countries than in middle- and low-income countries.

1.5 Main Publications

This cumulative thesis consists of the following publications:

- **Article 1:** Posch, L., Bleier, A., Schaer, P., and Strohmaier, M. (2015). "The Polylingual Labeled Topic Model." In: *KI 2015: Advances in Artificial Intelligence*.
- **Article 2:** Stier, S., Posch, L., Bleier, A., and Strohmaier, M. (2017). "When populists become popular: Comparing Facebook use by the right-wing movement Pegida and German political parties." In: *Information, Communication & Society 20.9*.

- **Article 3:** Eberhard, L., Walk, S., Posch, L., and Helic, D. (2019). "Evaluating narrative-driven movie recommendations on Reddit." In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*.
- **Article 4:** Posch, L., Bleier, A., Flöck, F., and Strohmaier, M. (2018). "Characterizing the global crowd workforce: A cross-country comparison of crowdworker demographics." *arXiv:1812.05948*.
- **Article 5:** Posch, L., Bleier, A., Lechner, C. M., Danner, D., Flöck, F., and Strohmaier, M. (2019). "Measuring motivations of crowdworkers: The Multidimensional Crowdworker Motivation Scale." In: *ACM Transactions on Social Computing 2.2*.

Furthermore, the following publications are related to the topics addressed in this thesis:

- Posch, L., Schaer, P., Bleier, A., and Strohmaier, M. (2016). "A system for probabilistic linking of thesauri and classification systems." In: *KI – Künstliche Intelligenz 30.2*.
- Posch, L., Panahiazar, M., Dumontier, M., and Gevaert, O. (2016). "Predicting structured metadata from unstructured metadata." In: *Database 2016*.
- Niekler, A., Bleier, A., Kahmann, C., Posch, L., Wiedemann, G., Erdogan, K., Heyer, G., and Strohmaier, M. (2018). "iLCM - A virtual research infrastructure for large-scale qualitative data." In: *Proceedings of the 11th International Conference on Language Resources and Evaluation*.
- Stier, S., Bleier, A., Bonart, M., Mörsheim, F., Bohlouli, M., Nizhegorodov, M., Posch, L., Maier, J., Rothmund, T., and Staab, S. (2018b). "Systematically monitoring social media: The case of the German federal election 2017." *arXiv:1804.02888*.

1.6 Contributions and Implications

This section gives an overview of the main contributions made in this thesis. The main contributions of this thesis are threefold and contribute to our understanding of *how* the microtask workforce is used in different stages of the machine learning process, *who* chooses to participate in it, and *why* people choose to participate in it. While the overarching research questions of this thesis are empirical in nature, the thesis makes both methodological and empirical contributions.

- First, this thesis demonstrates *how* microtasks can be employed in different stages of the machine learning process, to complement automated methods for the analysis of large text corpora. The usefulness of the microtask workforce for the analysis of large text corpora is demonstrated in three use cases, covering datasets from three different domains. In each of the use cases presented in this thesis, we employed microtasks in a different stage of the machine learning process. The use cases presented in this thesis additionally contain separate, both methodological and empirical, contributions, addressing the specific research questions posed by the individual projects that employed the microtask workforce. In each case, crowdworkers contributed human input that was indispensable for answering the project's research questions.
- Second, this thesis presents the first large-scale, country-level analysis of *who* participates in the international microtask workforce. It provides an analysis and comparison of the workforce's socio-demographic characteristics in ten different countries, a comparison of the importance of microtasks in the workers' lives, as well as an analysis of the country-level stability of these characteristics. This analysis advances our understanding of the composition of the international microtask workforce and provides important insights for researchers and policy makers seeking to understand this new form of work.

- Third, this thesis presents the first internationally applicable instrument for measuring *why* people choose to participate in the microtask workforce, and it presents the first validated cross-country comparison of workers' motivations. The measurement instrument for crowdworker motivations presented in this thesis enables researchers to incorporate workers' motivations in future studies investigating the microtask workplace. The results of the cross-country comparison of motivations provide important insights into the motivations of the international microtask workforce and shed light on the similarities and differences between the workers' motivations in different countries.

The contributions made in this thesis constitute an important step towards a more comprehensive characterization of the international microtask workforce and advance our understanding of the microtask workplace. The results of the analyses and the instrument for measuring crowdworker motivations provide a basis for future research concerning this emerging form of work and can help to inform policy discussions on where in the employment space microtasks should be located. For microtask platform providers, knowledge of the socio-demographic characteristics and motivations of their workforce can help to inform future design choices. Furthermore, the use cases presented in this thesis are relevant for data scientists concerned with the analysis of large text corpora as they provide insights into how the microtask workforce can be employed in different stages of the machine learning process to complement automated methods.

1.7 Structure of this Thesis

The remainder of this thesis is structured as follows. Chapter 2 gives an overview of related work that is relevant to the topics addressed in this thesis. Section 2.1 introduces the main methods used throughout this thesis, Section 2.2 introduces self-determination theory, the theory of motivation used in this thesis, and Section 2.3 gives an overview of related work regarding the use, socio-demographic characteristics, and motivations of the microtask workforce.

Chapter 3 presents the main publications contained in this cumulative thesis. First, Section 3.1 details my contributions to the individual publications. Section 3.2 focuses on the use of the microtask workforce in different stages of the machine learning process. It presents three use cases, in each of which microtasks complemented automated methods in a different stage of the machine learning process. Section 3.3 focuses on the characteristics of the international microtask workforce. It first presents an analysis of the socio-demographic characteristics of the microtask workforce in Section 3.3.1, and Section 3.3.2 presents the development and validation of the Multidimensional Crowdworker Motivation Scale as well as a cross-country comparison of crowdworker motivations.

Chapter 4 concludes this thesis. In Section 4.1, it first summarizes the main results and contributions of this thesis. Section 4.2 then discusses the implications of this work and describes a number of potential applications. Finally, Section 4.3 discusses the limitations of this thesis and outlines how these limitations open up directions for future work.

Figure 1.2 provides a structural overview of the topics addressed in this thesis and illustrates in which sections of this thesis the different topics are addressed. The machine learning process depicted in the figure represents a high-level view of the process and is similar to existing representations (see, e.g., Fayyad et al., 1996; Amershi et al., 2019).

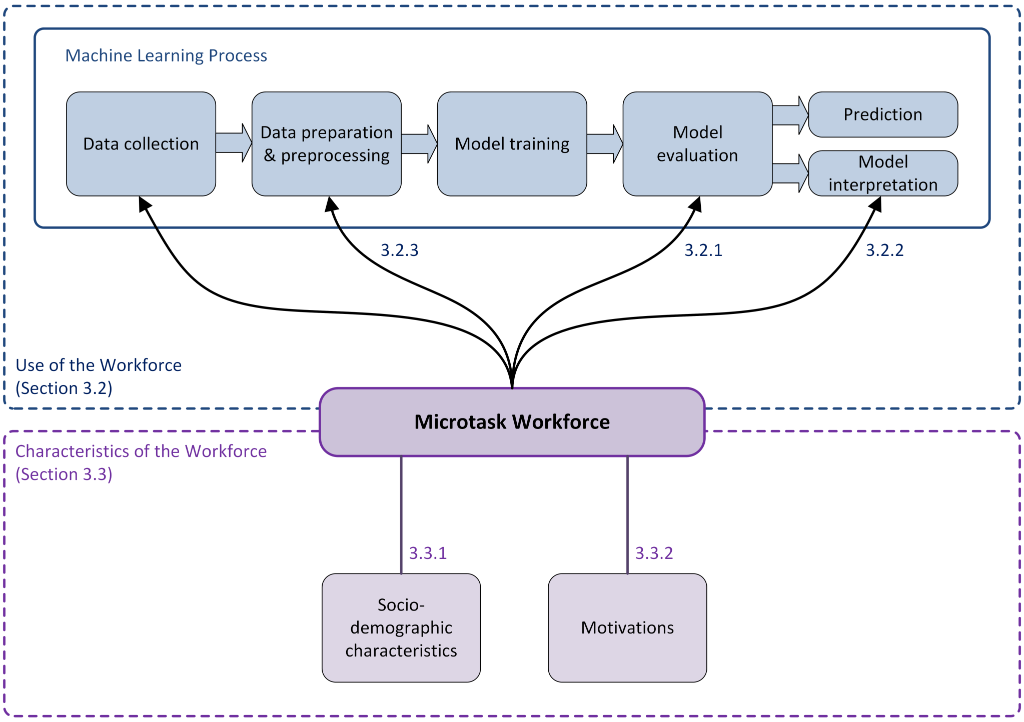


Figure 1.2: **Structural Overview of this Thesis.** This figure provides an overview of the topics addressed in this thesis. Section 3.2 focuses on the use of the microtask workforce in different stages of the machine learning process. The presented use cases employed microtasks in the *model evaluation* stage (Section 3.2.1), the *model interpretation* stage (Section 3.2.2) and the *data preparation and preprocessing* stage (Section 3.2.3) of the machine learning process. Additionally, we employed microtasks in the *data collection* stage in Stier, Bleier, Bonart, Mörsheim, et al. (2018b). Section 3.3 focuses on the characteristics of the international microtask workforce, presenting analyses of the socio-demographic characteristics (Section 3.3.1) and motivations (Section 3.3.2) of the workforce.

2 Related Work

This chapter introduces the main methods used throughout this thesis and gives an overview of related work. The chapter begins with an introduction to different methods for the measurement of latent variables in structured as well as unstructured data. Then, in Section 2.2, the central theory for motivation used in this thesis, self-determination theory, is introduced. Finally, Section 2.3 gives an overview of related work regarding the use, socio-demographic characteristics, and motivations of the microtask workforce.

2.1 Methods for the Measurement of Latent Variables

This section introduces the main methods used in this thesis to measure the concepts of interest in different contexts. In many contexts, it is often of central interest to measure phenomena that are not directly measurable or directly observable. For example, the topics addressed in a text document or theoretical concepts such as intelligence, work motivation, or populism cannot be directly observed.

Such unobservable and not directly measurable concepts are termed *latent constructs* or *latent variables* (see, e.g., Kline, 2015; Hair et al., 2018). While latent variables are unobservable, they can be represented by observable variables. By examining the observable variables that represent the latent variables, the latent variables can be measured indirectly. For example, latent topics occurring in a text document can be represented by a probability distribution over observable words, and latent constructs in survey response data can be represented by the observable survey item responses. The observable variables can be obtained from various data sources and data collection methods, such as surveys, observational methods, social media websites, or phone call records (Hair et al., 2018).

The methods introduced in this section all allow for measuring latent variables in observed data. Depending on the nature of the data, different methods for the analysis of latent variables are applicable, and this section introduces methods for both structured and unstructured data. Following the common distinction between structured and unstructured data (see, e.g., Rusu et al., 2013; Sint et al., 2009; Weglarz, 2004), I define *structured data* as data that follows a specific, predefined data model, i.e., the type of data that can be stored in a relational database. Analogously, I define *unstructured data* as any data that does not follow a predefined data model, such as the unstructured text contained in a collection of text documents.

The first part of this section introduces topic models, the main method used in this thesis for the measurement of latent variables in unstructured text data. The second part of this section introduces two methods for measuring latent variables in structured data, exploratory factor analysis and confirmatory factor analysis.

2.1.1 Measuring Latent Topics in Unstructured Text Data

A text document often addresses multiple topics. For example, one text document might address the topic of automated text analysis, the topic of microtasks, and the topic of populism in political communication. Another text document might address the topic of work motivation, the topic of microtasks, and the topic of scale validation.

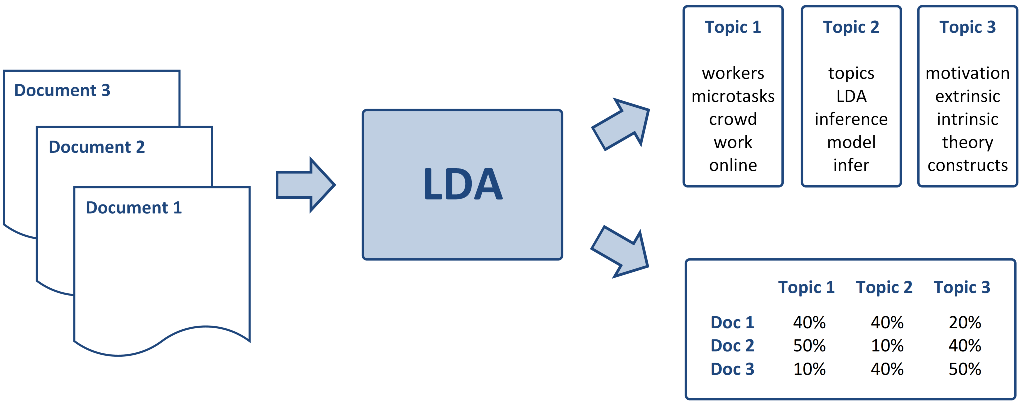
To understand the content of a document, a first step is to identify which topics the document addresses (T. L. Griffiths and Steyvers, 2004). However, in a collection of unstructured text documents, only the words themselves are observable variables. The topics that a document addresses are not directly observable; they are latent variables.

To infer these latent variables, a class of statistical models has been developed that represents the semantic properties of words and documents in terms of probabilities (see, e.g., Blei, Ng, et al., 2003; Heinrich, 2005; Steyvers and T. Griffiths, 2007; Blei, 2012; Barber, 2012). In these statistical models, called *topic models*, topics are represented as probability distributions over words and documents are represented as mixtures of topics.

Topic models are *generative models*, which means that the model specifies a procedure, called *generative storyline*, by which documents are generated (Steyvers and T. Griffiths, 2007). In the generative storyline of a topic model, a document is generated by first choosing a distribution over topics that will occur in the document. Then, to generate each word in the document, a topic is drawn from the document's topic distribution, and the word to generate is drawn from that topic's distribution over words.

When a topic model is trained on a corpus of existing documents, this generative storyline is inverted, and a set of topics that were responsible for generating the corpus is inferred by statistical techniques (Steyvers and T. Griffiths, 2007; Blei, 2012). In other words, the goal of inference in topic models is to estimate the model's parameters so that the identified latent variables (i.e., the topics) explain the observed variables (i.e., the words).

Figure 2.1 shows the general functionality of a typical topic model. A topic model takes a collection of text documents as input. Training the topic model means inferring the topics' word distributions and the documents' topic distributions. The result of the trained model is therefore a number of topics, represented by probability distributions over words, as well as a probability distribution over topics for each text document in the corpus.



	Topic 1	Topic 2	Topic 3
Doc 1	40%	40%	20%
Doc 2	50%	10%	40%
Doc 3	10%	40%	50%

Figure 2.1: **A Topic Model.** This figure shows the general functionality of a topic model. The model takes a collection of documents as input and estimates the word distributions of the topics as well as the documents' topic distributions.

Latent Dirichlet Allocation

Blei, Ng, et al. (2002) introduced one of the most widely used topic models, latent Dirichlet allocation (LDA). LDA is a generative Bayesian model that places Dirichlet priors on the topic distributions of the documents and on the word distributions of the topics. The model follows a *mixed membership* assumption, meaning that each document is not only described by a single latent topic but modeled as a mixture of multiple topics (see, e.g., Barber, 2012).

In LDA, a document d is a vector of N_d words, $\boldsymbol{w}_d$, where each word w_{di} is chosen from a vocabulary of V terms. A collection of documents is defined by $\mathcal{D} = \{\boldsymbol{w}_1,...,\boldsymbol{w}_D\}$. The number of topics K has to be specified a priori.[10] The generative storyline of LDA consists of the following steps:

[10] In parametric models such as LDA, the number of topics has to be specified a priori. There also exist non-parametric models, where the required number of topics is estimated during inference (see, e.g., Teh, Jordan, et al., 2006).

1. In the first step, for each document $d \in \{1,...,D\}$, a distribution θ_d over topics is drawn from a symmetric K-dimensional Dirichlet prior parametrized by α, representing the prior observation counts:

$$\theta_d \sim Dir(\alpha) \ . \tag{2.1}$$

2. In the second step, for each topic $k = \{1,...,K\}$, a distribution ϕ_k over the vocabulary is drawn from a V-dimensional Dirichlet distribution parametrized by β:

$$\phi_k \sim Dir(\beta) \ . \tag{2.2}$$

3. Finally, the i^{th} word in document d is generated by first drawing a topic index z_{di} and subsequently, a word w_{di} from the topic indexed by z_{di}:

$$w_{di} \sim Cat(\phi_{z_{di}}) \ , \qquad\qquad z_{di} \sim Cat(\theta_d) \ . \tag{2.3}$$

Based on this generative storyline, a number of inference techniques have been developed for reversing the storyline and estimating the latent variables that best explain the observed variables, i.e., the words in the documents. Commonly used inference techniques include Gibbs sampling (S. Geman and D. Geman, 1984), which is a type of Markov chain Monte Carlo inference, and collapsed variational Bayesian inference (Teh, Newman, et al., 2006).

Variations of Latent Dirichlet Allocation

Since the introduction of LDA, the model has been adapted to a variety of specific problem settings. In the following, I provide an overview of the adaptations most relevant to the problems addressed in this thesis.

Ramage et al. (2009) introduced *Labeled LDA (L-LDA)*, a supervised version of LDA. L-LDA is used to model the topics in a corpus of text documents where each document is annotated with multiple labels. A document's labels indicate which topics are present in the document, and L-LDA creates a topic for each label. Therefore, in L-LDA, a document d's topic distribution θ_d is restricted to a subset of all possible topics $\mathbf{\Lambda}_d \subseteq \{1,..,K\}$ – those topics with which the document is labeled. A collection of documents is then defined by $\mathcal{D} = \{(\boldsymbol{w}_1,\mathbf{\Lambda}_1),...,(\boldsymbol{w}_D,\mathbf{\Lambda}_D)\}$.

The first step in L-LDA's generative storyline draws the distribution of topics θ_d for each document $d \in \{1,...,D\}$

$$\theta_d \sim Dir(\alpha \boldsymbol{\mu}_d) \, , \tag{2.4}$$

where α is a continuous positive valued scalar representing the prior observation counts and $\boldsymbol{\mu}_d$ is a K-dimensional vector

$$\mu_{dk} = \begin{cases} 1 & \text{if } k \in \mathbf{\Lambda}_d \\ 0 & \text{otherwise} \, , \end{cases} \tag{2.5}$$

indicating which topics are permitted in each document according to the document's labels. Once these label-restricted topic distributions are drawn, the process of generating documents continues identically to the generative process of LDA shown above in Equations 2.2 and 2.3. In the case of $\mathbf{\Lambda}_d = \{1,..,K\}$ for all documents, no restrictions are active, and L-LDA is equivalent to LDA.

The generative view of LDA was extended to multilingual documents by Ni et al. (2009). Elaborating on this concept, Mimno et al. (2009) then developed the *Polylingual Topic Model (PLTM)*. This model assumes that the documents in the corpus are available in L different languages and that each of the L languages has a separate vocabulary. Each topic in the PLTM is therefore multilingual, having a word distribution in each language. In the special case of the documents being present in just one language, i.e., $L = 1$, the PLTM is reduced to LDA.

In the PLTM, a document d is represented by $[\boldsymbol{w}_d^1,...,\boldsymbol{w}_d^L]$, where for each language $l \in 1,...,L$, the vector $\boldsymbol{w}_d^l$ consists of N_d^l words which are chosen from a language-specific vocabulary with V^l terms. A collection of documents is then defined by $\mathcal{D} = \{[\boldsymbol{w}_1^1, ..., \boldsymbol{w}_1^L], ..., [\boldsymbol{w}_D^1, ..., \boldsymbol{w}_D^L]\}$.

The generative storyline of the PLTM consists of the following steps[11]:

1. The first step is identical to LDA: For each document $d \in \{1,...,D\}$, a distribution θ_d over topics is drawn from a symmetric K-dimensional Dirichlet prior parametrized by α:

$$\theta_d \sim Dir(\alpha) \ . \tag{2.6}$$

2. In the second step, for each topic $k = \{1,...,K\}$ in each language $l \in \{1,...,L\}$, a language-specific distribution ϕ_k^l over the language-specific vocabulary of length V^l is drawn:

$$\phi_k^l \sim Dir(\beta^l) \ . \tag{2.7}$$

3. Finally, the i^{th} word of language l in document d is generated by drawing a topic index z_{di}^l and subsequently, a word w_{di}^l from the language-specific distribution indexed by z_{di}^l:

$$w_{di}^l \sim Cat(\phi_{z_{di}^l}^l) \ , \qquad\qquad z_{di}^l \sim Cat(\theta_d) \ . \tag{2.8}$$

Apart from L-LDA and the PLTM, other variations of LDA have been developed for specific problem settings and to include additional information from the documents. Examples of adaptations for other specific settings are the *Author-Topic Model* (Rosen-Zvi et al., 2004; Bleier, 2012) that includes information about the authorship of documents, *Topics over Time* (Wang and McCallum, 2006), which jointly models word co-occurrences and localization in time, and the *Citation Influence Model* (Dietz et al., 2007) that includes citation information and estimates the strength of influence that one publication has over another.

[11] Note that the generative storyline of the PLTM is equivalent to LDA's except that steps 2 and 3 are repeated for each language.

In Section 3.2.1, this thesis extends existing work by introducing a new topic model, the *Polylingual Labeled Topic Model (PLL-TM)*, that combines the functionality of L-LDA and the PLTM. We applied the PLL-TM to a corpus of documents from the social science domain to measure the presence of social science concepts in the documents and to create probabilistic links between the concepts in a thesaurus and the concepts in a classification system from the same domain. The microtask workforce evaluated the semantic coherence of the topics produced by the PLL-TM, the PLTM, L-LDA and LDA. Furthermore, in Section 3.2.2, we employ LDA to model communication by German political parties on social media, with the aim of measuring the construct of populism. The microtask workforce interpreted the resulting model in the context of populist communication.

2.1.2 Measuring Latent Constructs in Structured Data

This section gives an overview of the methods used in this thesis for measuring latent constructs in structured data. The first part of this section introduces exploratory factor analysis (see, e.g., Harman, 1976; Thompson, 2004; Costello and J. Osborne, 2005; Hair et al., 2018), a method for identifying the structure underlying a set of observable variables. The second part of this section introduces confirmatory factor analysis (see, e.g., Bollen, 1989; D. Kaplan, 2008; Ullman and Bentler, 2003; Kline, 2015; Hair et al., 2018), a method for testing how well latent constructs, specified according to theory, represent the empirical data.

While both methods are used to measure latent constructs in structured data, they differ in their fundamental approach. Exploratory factor analysis does not require the latent variables and their relationships to the observable variables to be specified in advance. In contrast, confirmatory factor analysis requires that a model, derived from theory, is specified first.

Exploratory Factor Analysis

Exploratory factor analysis (EFA) is a multivariate data analysis technique for structured data that is based on the common factor model (see, e.g., Thurstone, 1947; Harman, 1976; Fabrigar et al., 1999). EFA uses the empirical data as a starting point and does not require any a priori hypotheses about the relationship between the observed variables and the latent variables, or about the number of latent variables. The method's primary purpose is to explore the data and to identify the structure underlying the observed variables (see, e.g., J. W. Osborne et al., 2008; Kline, 2015; Hair et al., 2018).

In EFA, *factors* are latent variables that are assumed to account for the correlations between the observed variables. The relations between the observed variables and the factors are termed *factor loadings* (see, e.g., Fabrigar et al., 1999; Hair et al., 2018). As EFA does not require the relationships between the latent variables and the observable variables to be specified in advance, it estimates loadings for all observable variables for each factor.

While all observed variables load on all factors, a factor structure emerges when observed variables have high loadings on a single factor and low loadings on all other factors (see, e.g., J. W. Osborne et al., 2008; Hair et al., 2018). Examining the factor loadings of the observed variables therefore gives insights into the nature of the factors and helps to identify a set of observed variables that are suitable to represent a latent construct (Hair et al., 2018).

As factor models that have more than one factor do not have a unique solution in EFA, the reference axes of the factors can be rotated in multidimensional space (Fabrigar et al., 1999; Jennrich, 2007). The goal of this rotation is simplifying the structure of the solution (see, e.g., Fabrigar et al., 1999; Jennrich, 2007; Hair et al., 2018). There are two general types of axis rotation methods: *Orthogonal rotation* adjusts the factor axes so that the factors are constrained to be independent of each other, and *oblique rotation* allows the factors to be correlated (Fabrigar et al., 1999; Jennrich, 2007; Hair et al., 2018).

EFA can help to identify which observable variables (e.g., survey items) are appropriate for representing a theoretical latent construct (see, e.g., Worthington and Whittaker, 2006). Performing EFA on a set of observable variables, the estimated factor loadings indicate which of the variables are likely to represent a construct well and which ones do not represent the construct they were intended to represent. Furthermore, if an observed variable has high loadings on multiple factors (termed *cross-loadings*), this indicates that the observed variable has a strong relationship with more than one latent variable and that it is therefore not suitable for unambiguously representing and measuring a single latent construct.

When a new instrument for measuring certain theoretical latent constructs is developed, EFA is often performed on a pool of observed candidate variables, with the goal of reducing and refining the pool by deleting variables that exhibit undesirable properties such as high loadings on multiple factors or no high loadings on any factor (Worthington and Whittaker, 2006). This process leads to retaining a set of observable variables that are likely to be suitable for representing and measuring the latent constructs. The suitability of this set of observable variables can then be further validated by performing confirmatory factor analysis.

Confirmatory Factor Analysis

Confirmatory factor analysis (CFA) is a multivariate data analysis technique used to test how well latent constructs that are specified according to theory represent reality according to the data gathered (Hair et al., 2018). In contrast to EFA, CFA requires the development and specification of a measurement theory. This a priori specification must include how many latent constructs exist in the model and how the observed variables correspond to the latent constructs (Kline, 2015; Hair et al., 2018).

CFA is part of a set of techniques called *structural equation modeling* (SEM) (see, e.g., Bollen, 1989; D. Kaplan, 2008; Ullman and Bentler, 2003; Hoyle, 2012; Kline, 2015; Hair et al., 2018). SEM aims to analyze structural relationships in multivariate data, on the basis of a theoretical model. The set of techniques enables the estimation of multiple interrelated dependence relationships while also allowing multiple measures for each latent construct (Hair et al., 2018). In SEM, the specification of the measurement theory is termed *measurement model*. CFA assesses this measurement model, i.e., it evaluates the hypothesized relationships between the observed variables and the latent constructs in the model by fitting the model to the observed data (see, e.g., Kline, 2015; Hair et al., 2018).

Structural equation models are visualized using *path diagrams*. Path diagrams represent the set of structural equations that specify the model. Figure 2.2 shows an exemplary path diagram for a CFA model with two latent constructs and six observed variables. In this model, each latent construct is measured by three observed variables.

In path diagrams, observed variables are represented by rectangles or squares, and latent variables are depicted as ellipses or circles. Error terms of observed variables, i.e., their variance that is not explained by the associated latent variables (Kline, 2015), are also represented by ellipses or circles.

Double-headed arrows indicate correlational relationships. In the model shown in Figure 2.2, the two latent constructs are hypothesized to correlate. Single-headed arrows between latent constructs and observed variables indicate hypothesized directional effects. In a CFA model, latent constructs have a presumed causal effect on their associated observed variables (Kline, 2015). For example, a person's performance on a certain task (an observable variable) may be presumed to be caused by an underlying latent construct (for example, "intelligence"). The single-headed arrow from the error terms to the observed variables represents the assumption that the observed variable is not only caused by the latent variable, but also by other, unmeasured causes (Kline, 2015).

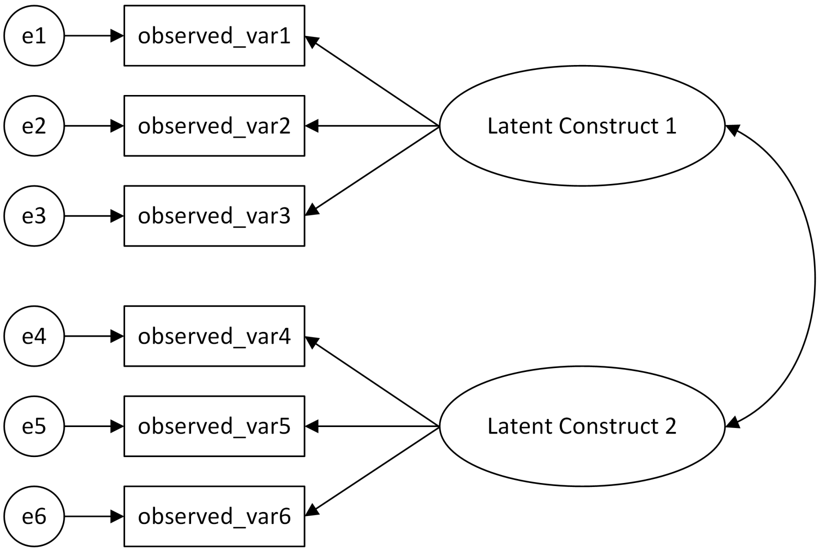


Figure 2.2: **A CFA Model.** This figure shows an exemplary CFA model with two latent constructs. In the model, each latent construct is measured by three observed variables, and the latent constructs are hypothesized to correlate.

Fitting the hypothesized model means estimating the model's parameters, including factor variances and covariances, factor loadings, and the amount of measurement error for each observed variable (see, e.g., Kline, 2015). Factor loadings, in the context of CFA, are estimates of the presumed causal effects that the latent constructs have on the observed variables. The most commonly used method for estimating the parameters of a structural equation model is maximum likelihood estimation (Jöreskog, 1970; Kline, 2015), but alternative estimation methods, such as general least squares (Jöreskog and Goldberger, 1972), also exist.

Evaluating Model Fit. The model fit of a CFA model can be evaluated via a range of goodness-of-fit (GOF) measures, which assess the extent to which the theory, as specified in the model, represents reality as observed in the data (see, e.g., McDonald and Ho, 2002; Sun, 2005; Marsh et al., 2005; Schreiber et al., 2006; Hooper et al., 2008; Kline, 2015; Hair et al., 2018). A basic test statistic to measure GOF is the *model chi-square (χ^2)*. The

χ^2 test, in the context of CFA, is a statistical test of the difference between the observed covariance matrix and the covariance matrix estimated by the model (see, e.g., Hooper et al., 2008; Kline, 2015). The null hypothesis is that the observed and estimated matrices are identical, i.e., it tests for exact fit.

The model χ^2 is an *absolute fit index*, which means that it measures how well the model represents the data, independently of alternative models (McDonald and Ho, 2002; Hooper et al., 2008; Hair et al., 2018). While χ^2 is commonly reported (Worthington and Whittaker, 2006), it should not necessarily be relied upon for determining model fit as it is sensitive to sample size and to the number of observed variables in the model (see, e.g., Bentler and Bonett, 1980; Miles and Shevlin, 2007; Hair et al., 2018).

Due to the limitations of the χ^2 test, a variety of alternative GOF measures have been developed to assess model fit. Alternative absolute fit indices include the *root mean squared error of approximation* (RMSEA), the *standardized root mean square residual* (SRMR), and the *Goodness-of-Fit statistic* (GFI). *Incremental fit indices* compare the specified model with a baseline model where all variables are uncorrelated (see, e.g., McDonald and Ho, 2002; Miles and Shevlin, 2007; Hair et al., 2018). Examples of incremental fit indices are the *normed-fit index* (NFI), the *comparative fit index* (CFI) and the *Tucker Lewis Index* (TLI). *Parsimony fit indices*, such as the *adjusted goodness of fit index* (AGFI) and the *parsimony normed fit index* (PNFI), are used for comparing competing models and favor simpler models over complex models (see, e.g., Kline, 2015; Hair et al., 2018).

There have been decades of discussion on which fit measures should be used to determine model fit and which cut-offs should be used to either accept or reject a model (see, e.g., Hooper et al., 2008). For current conventions, see, for example, Kline (2015), Sun (2005), Hooper et al. (2008), and Schreiber et al. (2006), who describe the different fit measures in detail and discuss which values indicate good or adequate model fit.

Measurement Invariance. Tests of *measurement invariance* assess *"whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute"* (Horn and McArdle, 1992). When interpreting differences in the means of latent constructs across groups, it is critical to first ensure that the instrument used to measure the latent constructs of interest is invariant across the groups (see, e.g., Vandenberg and Lance, 2000; Cheung and Rensvold, 2000; Millsap and Olivera-Aguilar, 2012; Millsap, 2012).

If measurement invariance is not established, any conclusions drawn from a comparison of latent construct means are necessarily ambiguous or even fallacious (Steenkamp and Baumgartner, 1998). For example, in a cross-country comparisons of survey results, a lack of measurement invariance could indicate that respondents of different countries understand the survey items in a different way and associate them with different latent constructs (e.g., due to culture), that the strength of these associations differs, or that different levels of response biases are present (see, e.g., Cheung and Rensvold, 2000).

Measurement invariance is commonly evaluated via multiple-group confirmatory factor analysis (MGCFA) (Jöreskog, 1971; Cheung and Rensvold, 2002), and three levels of measurement invariance are commonly tested: configural, metric, and scalar invariance (see, e.g., Cheung and Rensvold, 2002; F. F. Chen, 2007; Millsap, 2012; Putnick and Bornstein, 2016). *Configural invariance* requires that the observed variables share the same configurations of factor loadings in all groups. *Metric invariance* additionally requires that the loadings of the observed variables on their factors are equal across groups, and *scalar invariance* additionally requires that the intercepts of the observed variables are the same across groups. Several studies have examined the sensitivity of GOF measures to a lack of measurement invariance. For example, Cheung and Rensvold (2002) and F. F. Chen (2007) conducted simulation studies and provided recommendations for evaluating the different levels of measurement invariance.

To validly compare observed composite means across groups, scalar invariance is required. For example, a simulation study by Steinmetz (2013) showed that even one unequal intercept may lead to erroneous conclusions about differences in means when comparing observed composite scores across groups. At least partial scalar invariance is required to validly compare model-estimated latent means across groups (see, e.g., Byrne et al., 1989; Vandenberg and Lance, 2000).

In Section 3.3.2, this thesis presents a new measurement instrument for measuring work motivation in the microtask context. We conducted EFAs on data collected in three countries to develop the set of observable variables to be used in the new measurement instrument. The evaluation of work motivation scales developed for the traditional work context and the validation of the new measurement instrument in ten countries and three income groups were conducted with CFAs. To establish cross-group comparability of latent construct means between countries and between country income groups, we tested the measurement invariance of our model.

2.2 Self-Determination Theory

Self-determination theory (SDT) is an empirically based theory of human motivation and personality development that was developed by Deci and Ryan (see, e.g., Deci and Ryan, 1980; Deci and Ryan, 1985; Ryan and Deci, 2000; Deci and Ryan, 2000; Deci and Ryan, 2002; Deci, Olafsen, et al., 2017; Ryan and Deci, 2017). SDT is comprised of a set of mini-theories, including *Cognitive Evaluation Theory*, *Organismic Integration Theory*, *Causality Orientations Theory*, *Goal Content Theory*, *Basic Psychological Needs Theory*, and *Relationships Motivation Theory* (see, e.g., Deci and Ryan, 2002; Ryan and Deci, 2017). Each of these mini-theories addresses a facet of human motivation.

SDT postulates that there are three basic innate psychological needs that are essential for an individual's psychological growth, integrity, and well-being: autonomy, competence, and relatedness (Deci and Ryan, 2000). *Autonomy* refers to the perception of being the origin of one's behavior, *competence* refers to the feeling of being effective and being able to express one's capacities, and *relatedness* refers to the feeling of being connected to others (Deci and Ryan, 2002; Deci and Ryan, 2000). Contexts that support the satisfaction of the innate psychological needs are associated with different types of motivation than contexts that prevent the satisfaction of these needs (Deci and Ryan, 2000).

In contrast to most other theories of motivation, which consider motivation to be a unitary concept and focus mainly on the total *amount* of motivation an individual has, SDT focuses on the *type* of motivation (Ryan and Deci, 2000; Gagné and Deci, 2005). The theory distinguishes between three general types of motivation: *intrinsic motivation*, *extrinsic motivation*, and *amotivation*. *Intrinsic motivation* is a non-instrumental type of motivation. When intrinsically motivated, people act freely and are driven by interest and enjoyment inherent in the action (Ryan and Deci, 2000). *Extrinsic motivation*, in contrast, is instrumental. When extrinsically motivated, an individual engages in an activity because it leads to an outcome that is separable from the activity itself (Ryan and Deci, 2000). In contrast to both intrinsic motivation and extrinsic motivation, *amotivation* is non-intentional. It is the absence of motivation, a state of acting passively or not intending to act all (Deci and Ryan, 2000).

In SDT, the different types of motivation are hypothesized to lie along a continuum of self-determination: At the one extreme of the continuum lies *amotivation*, which is completely lacking in self-determination; at the other extreme lies *intrinsic motivation*, which is completely self-determined (Gagné and Deci, 2005). *Extrinsic motivation*, which lies between these extremes, is further split up into subtypes that differ in the degree to which they are autonomous: *external regulation*, *introjected regulation*, *identified regulation*, and *integrated regulation* (see, e.g., Ryan and Deci, 2000; Deci and Ryan, 2000; Deci and Ryan, 2002; Ryan and Deci, 2017).

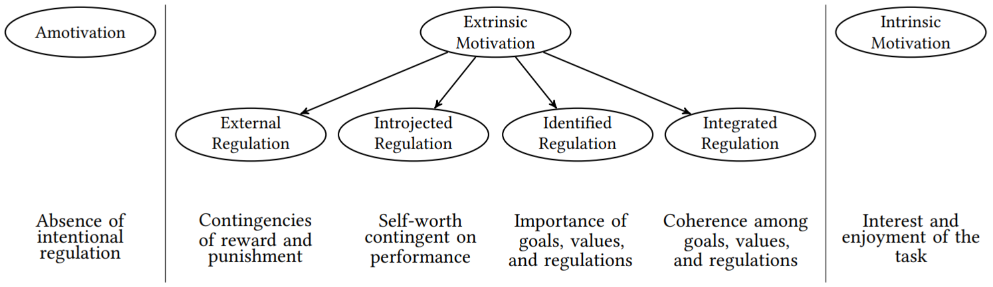

Amotivation
Extrinsic Motivation
Intrinsic Motivation
External Regulation
Introjected Regulation
Identified Regulation
Integrated Regulation
Absence of intentional regulation
Contingencies of reward and punishment
Self-worth contingent on performance
Importance of goals, values, and regulations
Coherence among goals, values, and regulations
Interest and enjoyment of the task

Figure 2.3: **Types of Motivation.** This figure shows the different types of motivation along the self-determination continuum hypothesized by SDT. The figure is based on Gagné and Deci (2005).

Of the subtypes of extrinsic motivation, *external regulation* is the least self-determined. Individuals motivated by external regulation act in order to obtain external rewards or avoid punishments. *Introjected regulation* is a form of partially internalized extrinsic motivation where an individual engages in an activity to avoid feelings of guilt or to attain feelings of worth. *Identified regulation* is a form of extrinsic motivation with a high degree of perceived autonomy, where an individual engages in an action because it is consciously valued and in alignment with the individual's personal goals. The most self-determined form of extrinsic motivation is *integrated regulation*. Integrated regulation stems from evaluated identifications that are congruent with self-endorsed values, goals, and needs (Ryan and Deci, 2000; Deci and Ryan, 2002). While integrated regulation is highly self-determined, it is still a form of extrinsic motivation as the activity is engaged in to achieve an outcome that is separate from the activity itself (Ryan and Deci, 2000). Figure 2.3, based on Gagné and Deci (2005), depicts the different types of motivation as specified by SDT.

SDT postulates that individuals may internalize an initially external regulation, which then becomes more self-determined (Deci and Ryan, 2000). Depending on the extent to which the individual has integrated it with his or her sense of self, initially external regulations may be internalized in different ways (Deci and Ryan, 2002). For example, an individual might not perceive an activity as enjoyable but engage in this activity because it leads to a tangible reward. The individual might later internalize this externally regulated behavior, for example because he or she starts to perceive it to be important for his or her personal goals and therefore starts to value it.

2.2.1 Self-Determination Theory and Work Motivation

Work motivation has been defined as *"a set of energetic forces that originate both within as well as beyond an individual's being, to initiate work-related behavior, and to determine its form, direction, intensity, and duration"* (Pinder, 2014). Early models of work motivation, such as expectancy-valence theories (see, e.g., Vroom, 1964, for an overview, also see Pinder, 2014), considered motivation to be a unitary concept and therefore focused mainly on the total *amount* of motivation (Gagné and Deci, 2005). Porter and Lawler (1968) developed a theory that distinguished between extrinsic and intrinsic motivation, but they assumed that extrinsic and intrinsic work motivation were additive, i.e., that together, they would yield a worker's level of total job satisfaction (Porter and Lawler, 1968; Gagné and Deci, 2005).

Other research (e.g., Deci, 1971), however, suggested that certain types of extrinsic rewards diminished intrinsic motivation while other types of extrinsic rewards seemed to enhance it. Cognitive evaluation theory was developed to explain this interaction between extrinsic and intrinsic motivation (Deci and Ryan, 1980). A meta-study of 128 studies, conducted by Deci, Koestner, et al. (1999), later corroborated that all contingent tangible rewards had a negative effect on intrinsic motivation, whereas positive feedback ("verbal rewards") enhanced self-reported interest and free-choice behavior. Moreover, as Deci, Koestner, et al. (1999) argued, the use of rewards in organizations was likely to be accompanied by other factors that had also been found to undermine intrinsic motivation, such as increased surveillance, evaluation, and competition (Deci and Ryan, 1985).

Building on and incorporating cognitive evaluation theory, SDT was developed to provide a broader framework for studying human motivation (Deci and Ryan, 1985; Deci and Ryan, 2000). Gagné and Deci (2005) later described the implications of SDT as a theory of work motivation, and Deci, Olafsen, et al. (2017) presented a general SDT model of work motivation.

Several instruments for measuring work motivation in the traditional employment context have been developed based on SDT. These instruments measure work motivation at the domain level of analysis, which means that they measure the general motivation to perform a job as opposed to specific tasks within a job.

Blais et al. (1993) developed a French SDT-based work motivation scale, which was translated into English by Tremblay et al. (2009). The resulting Work Extrinsic and Intrinsic Motivation Scale (WEIMS) was evaluated in different work environments and measures six types of motivation: amotivation, four subtypes of extrinsic motivation (external regulation, introjected regulation, identified regulation, and integrated regulation), and intrinsic motivation. Gagné, Forest, Gilbert, et al. (2010) developed the Motivation at Work Scale (MAWS), an instrument that measures external regulation, introjected regulation, identified regulation, and intrinsic motivation. The MAWS was partly based on the scale developed by Blais et al. (1993), and it was validated in French and in English.

Gagné, Forest, Vansteenkiste, et al. (2015) later developed the Multidimensional Work Motivation Scale (MWMS), which does not include any items from the MAWS. The MWMS was validated in seven languages and nine countries, and it measures six first-order factors (amotivation, material external regulation, social external regulation, introjected regulation, identified regulation, and intrinsic motivation) and one second-order factor (external regulation).

In Section 3.3.2, this thesis builds on existing work by presenting the development and validation of the Multidimensional Crowdworker Motivation Scale (MCMS), an SDT-based instrument for measuring work motivation in the microtask context. The MCMS is the first SDT-based instrument that was developed specifically for this context and that provides a comprehensive representation of the motivational dimensions according to SDT.

2.3 The Microtask Workforce

This section gives an overview of related studies that investigated different aspects of the microtask workforce. First, Section 2.3.1 summarizes research regarding different uses of the microtask workforce, focusing on applications related to natural language processing. Section 2.3.2 then gives an overview of related work regarding the socio-demographic characteristics of the workforce. Finally, Section 2.3.3 describes existing studies on the motivations of the microtask workforce.

2.3.1 Use of the Workforce

This section summarizes research concerning different uses of the microtask workforce. The summary given in this section is not intended to be an exhaustive review of studies that have utilized the microtask workforce, but rather intends to give a general overview, with a focus on tasks related to natural language processing.

Several taxonomies of crowdsourcing have been created, both for crowdsourcing systems in general and for microtasks in particular. Yuen et al. (2011) presented a taxonomy of crowdsourcing based on a literature survey of crowdsourcing systems. They grouped crowdsourcing applications into four different general categories, including voting systems, information sharing systems, social games, and creative systems. Geiger, Seedorf, et al. (2011) created a taxonomy of crowdsourcing processes, identifying 19 different process types that they grouped into five clusters. Microtasks were located in the cluster “integrative sourcing with fixed remuneration," meaning that the input from the crowd is pooled, and all contributions are rewarded with a fixed payment. Geiger, Rosemann, et al. (2012) created a typology of crowdsourcing information systems, distinguishing between crowd rating systems, crowd creation systems, crowd processing systems, and crowd solving systems.

Gadiraju, Kawase, et al. (2014) created a taxonomy of typical microtasks, based on a survey of 490 workers on CrowdFlower. They asked workers open-ended questions about two tasks that they had recently completed and, based on the workers' responses, manually identified six different high-level types of microtasks as well as a number of sub-types of tasks. Their taxonomy distinguishes between the following high-level types of microtasks: information finding (e.g., metadata finding), verification and validation (e.g., spam detection), interpretation and analysis (e.g., classification), content creation (e.g., media transcription), surveys (e.g., feedback/opinions), and content access (e.g., promoting). Demartini, D. E. Difallah, Gadiraju, et al. (2017) provided an overview of how microtasks have been used in different hybrid human-machine information systems.

Use of Microtasks for Natural Language Processing. Processing and understanding natural language is typically easier for humans than for computers because the unstructured nature and ambiguity of natural language often pose a significant challenge to automated methods in many natural language processing (NLP) tasks. In contrast, humans generally understand natural language very well and are often intuitively capable of performing tasks such as identifying the context that a word is used in, identifying the sentiment of a statement, or identifying the most relevant information in a paragraph of text. For this reason, microtasks have the potential to support numerous NLP tasks. The remainder of this section gives an overview of how microtasks have been employed for complementing automated methods for text-related NLP tasks.[12]

[12] While microtasks have also been employed to support natural language processing tasks concerning the analysis of spoken language (see, e.g., Shashidhar et al., 2015;

Supervised text classification is a natural language processing application that requires labeled text data to train and evaluate the machine learning model. Labels for these training and test datasets typically require human annotators, and microtasks have proven to be a cost-effective and scalable method to create labeled text data (see, e.g., Snow et al., 2008; Hoffmann, 2009). There has been a significant amount of research on different aspects of employing microtasks for the creation of labeled datasets for supervised classification. For example, studies have investigated aspects such as the quality of annotations (e.g., Snow et al., 2008; Hsueh et al., 2009; Ipeirotis et al., 2010; Bu et al., 2019), methods of aggregating the answers of individual workers (e.g., Raykar, Yu, et al., 2010; Raykar and Yu, 2012; Venanzi et al., 2014; Simpson et al., 2015), the necessary number of workers (e.g., Karger et al., 2011), or the effects of training workers for a task (e.g., Gadiraju, Fetahu, et al., 2015). The application of microtasks for annotating pieces of content with objective labels has also been termed "crowdcoding" (Haselmayer and Jenny, 2017; Guo et al., 2019).

An NLP application that is closely related to text classification is sentiment analysis, which aims at identifying sentiments or attitudes expressed in unstructured text. If sentiment analysis is conducted with a supervised method, it is a type of text classification and requires training and test datasets in which the text is labeled with the sentiment expressed in it. Crowdworkers have been used to support supervised sentiment analysis by creating such labels in several studies.

For example, Hsueh et al. (2009) used microtasks to annotate text segments from political blogs with the sentiment expressed in them. Gadiraju, Fetahu, et al. (2015) employed crowdworkers for assessing the sentiment expressed in tweets and showed that training the workers for this task increased their accuracy. Borromeo and Toyama (2015) compared sentiment annotations created by crowdworkers with annotations created by a supervised classifier trained on movie reviews and found that, compared to the automatic method, the annotations created by crowdworkers

Caines et al., 2016), a review of this application of microtasks is outside the scope of this section.

had a higher agreement with their gold standard. Simpson et al. (2015) developed a Bayesian approach that combines crowdsourced annotations with text features to identify the sentiment of text documents.

Microtasks have also been used for annotating unstructured text with latent constructs for the purpose of quantitative content analysis. For example, Lind et al. (2017) investigated the potential of microtasks for annotating news texts with the latent construct of political actor evaluations. In this study, the annotation of text with a latent construct constituted a form of sentiment analysis, where the goal was to identify the sentiment concerning a specific target at the sentence level. The authors concluded that crowdworkers can be a reliable and valid alternative to expert coders for annotating text with latent constructs. Benoit et al. (2016) used microtasks for annotating a corpus of political texts along two policy dimensions.

Named entity recognition (NER) is another NLP-related task where microtasks have been shown to be complementary to automated methods. NER is concerned with identifying entities in natural language text, such as people, locations, products, or organizations. For example, Finin et al. (2010) used microtasks for identifying named entities in Twitter status updates, identifying persons, organizations, and locations, and Lawson et al. (2010) used microtasks for identifying named entities in emails. Feyisetan, Luczak-Roesch, et al. (2015) investigated how different features of microposts and crowdworker preferences were related to the accuracy and speed of crowdsourced named entity recognition. They found that crowdworkers performed well in identifying people, locations, and implicitly defined entities in short texts.

Related to NER is the task of entity linking, which aims at creating links between the entities mentioned in a text and the corresponding entities in a knowledge base (see, e.g., Shen et al., 2014). Demartini, D. E. Difallah, and Cudré-Mauroux (2012) showed that microtasks can be used to improve the quality of such links by combining algorithmic results with the results generated by crowdworkers. Bontcheva et al.

(2017) created a corpus of tweets in which named entities were annotated and linked to the knowledge base DBPedia by crowdworkers.

While many of the applications that used microtasks for NLP-related tasks constituted some type of annotation of unstructured text, often for the purpose of training and evaluating supervised classifiers, microtasks have also been employed for complementing unsupervised methods for the analysis of text. For example, Chang et al. (2009) proposed quantitative methods for evaluating different aspects of the interpretability of a topic model. Specifically, a "word intrusion" task was proposed to measure the semantic coherence of topics, and a "topic intrusion" task was proposed to measure the extent to which the estimated mixture of topics for a document corresponded to human perceptions of the document's content. Towne et al. (2016) used microtasks for evaluating topic models with respect to their ability to capture document similarity.

Other applications related to NLP for which the support of crowdworkers has been shown to be helpful include tasks such as detecting plagiarism (e.g., Potthast et al., 2010), translation (e.g., Callison-Burch, 2009; Pavlick et al., 2014; Yan et al., 2014), text editing (e.g., Bernstein et al., 2010), and word sense disambiguation (e.g., Parent and Eskenazi, 2010).

Section 3.2 of this thesis focuses on the use of the microtask workforce in different stages of the machine learning process. In three projects from different domains, this thesis employs microtasks to complement automated methods for text analysis. Each of these projects employed microtasks in a different stage of the machine learning process (also see Figure 1.2 in Section 1.7), and in each project, microtasks were essential for answering the respective project-specific research questions.

Specifically, in Section 3.2.1, we employed microtasks in the *model evaluation* stage. Following the method proposed by Chang et al. (2009), we compared the semantic coherence of three existing topic models with our newly proposed topic model for the task of modeling a corpus of documents from the social science domain. Section 3.2.2 presents an analysis of populist political communication, where we employed microtasks in the *model interpretation* stage. In Stier, Posch, et al. (2017), we modeled the communication by German political parties on social media over a time span of around 20 months, with the aim of measuring the use of populist communication by different parties. Crowdworkers interpreted the model parameters, i.e. the estimated topics, in the context of populist communication. While this constitutes a form of coding latent constructs, in contrast to Lind et al. (2017), we did not employ microtasks to directly annotate the unstructured text of the corpus, but rather to interpret latent topics estimated by an unsupervised machine learning model. Section 3.2.3 presents an evaluation of the capability of established recommender algorithms to incorporate information from narrative descriptions of users' preferences. We employed a range of microtasks in the *data preparation and preprocessing* stage of the machine learning process to extract structured data and from the unstructured text contained in posts on the platform Reddit. The microtasks included named entity recognition, sentiment analysis, and the extraction of other relevant information from the unstructured text. Additionally, in Stier, Bleier, Bonart, Mörsheim, et al. (2018b), we employed microtasks in the *data collection* phase for creating a collection of social media accounts owned by mainstream and alternative German media outlets.

2.3.2 Socio-Demographic Characteristics of the Workforce

Most studies investigating the socio-demographic characteristics of the microtask workforce have focused on the platform Amazon Mechanical Turk. The first part of this section gives an overview of these studies investigating MTurk's workforce. The second part of this section gives an overview of research regarding the socio-demographic characteristics of workers on the platform Figure Eight, the second market leader in the microtask market, and the third part of this section reviews research on the characteristics of workers on other microtask platforms.

Workers on Amazon Mechanical Turk. Since the launch of Amazon Mechanical Turk (MTurk) in 2005 (Amazon Mechanical Turk, 2015), various studies have been conducted that investigated the socio-demographic characteristics of this platform's anonymous workforce. Ross et al. (2010) and Ipeirotis (2010a) were among the first researchers to study the demographics of workers on MTurk, collecting responses from 573 and 1,000 workers, respectively. Both studies found that the vast majority of workers were located in the USA and India, with workers from the USA constituting the largest part of MTurk's workforce. Furthermore, the studies found that MTurk's workforce was diverse with regards to age, gender, and income, but that it was younger and more highly educated than the general population. Workers in the USA were predominantly female, while workers in India were predominantly male. A significant minority of workers in both studies reported that they worked on MTurk to be able to pay for basic expenses.

Similar results were reported by later studies on the socio-demographic characteristics of MTurk workers (see, e.g., Pavlick et al., 2014; Berg, 2015; Peer et al., 2017; Goodman and Paolacci, 2017; Naderi, 2018; D. Difallah et al., 2018). For example, Berg (2015) found that Indian and American workers on MTurk were young and highly educated and that the majority of workers in India were male, but that, in contrast to the findings of earlier studies, there was now gender balance among workers from the U.S. Futhermore, Berg (2015) found that, while many workers in the survey complemented their income from MTurk with income from other jobs, 49% of Indian workers and 38% of American workers reported MTurk as their primary source of income.

In addition to these studies, *mturk tracker*[13], a tool developed by Ipeirotis (2010b), allows for tracking the location, gender, age, marital status, household size, and household income of the MTurk workforce by posting a survey task to MTurk every 15 minutes (also see D. Difallah et al., 2018). According to mturk tracker, workers from the United States and India currently still constitute over 80% of the worker population on

[13] `https://www.mturk-tracker.com`

MTurk, which is likely due to the fact that workers located in other countries can only receive payment for their work in the form of Amazon.com gift cards (Amazon Mechanical Turk, 2018).

A number of studies have also investigated the representativeness of MTurk samples and the suitability of such samples for different research purposes (see, e.g., Paolacci, J. Chandler, and Ipeirotis, 2010; Buhrmester et al., 2011; Berinsky et al., 2012; Shapiro et al., 2013; Weinberg et al., 2014; Paolacci and J. Chandler, 2014; Huff and Tingley, 2015). An early study on the representativeness of MTurk's workforce, conducted by Paolacci, J. Chandler, and Ipeirotis (2010), found that the population of U.S. workers on MTurk was not less representative of the U.S. population than traditional university subject pools. Buhrmester et al. (2011) compared the demographics of a sample of MTurk workers to a large internet sample and concluded that their MTurk sample was more diverse than both standard internet samples and American college samples.

A study by Berinsky et al. (2012) evaluated the suitability of MTurk samples for experimental political science. They found that their sample of workers on MTurk was more representative of the U.S. population than in-person convenience samples, but less representative than respondents recruited for internet-based panels or national probability samples. Furthermore, they found that the way workers on MTurk responded to experimental stimuli was consistent with prior research. Paolacci and J. Chandler (2014) analyzed the suitability of MTurk workers as a participant pool for the social sciences and concluded that worker samples from MTurk should not be considered representative of a country's population, but could nevertheless replace or supplement convenience samples in psychological research.

Weinberg et al. (2014) compared the socio-demographic characteristics of workers on MTurk to those of respondents of a population-based web panel. In their study, the sample of MTurk workers was more divergent from the general population than the web panel. In the MTurk sample, the proportion of women was higher, and MTurk workers were younger and more educated than participants from the web panel. Huff and

Tingley (2015) compared the demographics and political characteristics of MTurk workers from the United States to the characteristics of the respondents of a stratified sample survey. They found that MTurk was good at attracting certain demographics that were difficult to attract for the stratified sample survey, and that the distribution of employment in different occupational sectors, as well as the location on the rural-urban continuum, was similar in both samples.

Workers on Figure Eight. Most research regarding the socio-demographic composition of the microtask workforce has focused on MTurk and therefore on workers located in the USA and India. Even though the microtask platform Figure Eight (formerly CrowdFlower) is the second market leader in the microtask market and has a revenue approximately equal to MTurk's (Kuek et al., 2015), its workforce has so far received surprisingly little attention in research. Furthermore, despite Figure Eight's workforce being much more international than MTurk's workforce, none of the studies concerned with Figure Eight's workforce have so far analyzed and compared the socio-demographic characteristics of the platform's workers at the country level.

Berg (2015) collected socio-demographic data from 353 workers on CrowdFlower and found that only 2.8% of the workers in the sample were located in the U.S. and 8.5% were located in India. Workers were predominantly male, and 31% of workers reported that the income from the platform was their primary source of income. Comparing the sample of workers on CrowdFlower to a sample of workers on MTurk, the study found that workers on CrowdFlower were more educated than American workers on MTurk, but less educated than Indian workers on MTurk. Peer et al. (2017) examined the demographics of workers on CrowdFlower ($N = 221$) and on the platform Prolific Academic ($N = 214$) and compared them to the demographics of workers on MTurk ($N = 201$). The study found that workers on all three platforms were highly educated and had a similar mean age. Compared to MTurk, the workers on CrowdFlower and Prolific Academic were much more geographically diverse, and both platforms had a higher proportion of male workers than MTurk. The study further examined whether workers tended to work on more than one platform rather than committing to a single platform, and the results indicated that the overlap between the workforces was small. Only 2.5% of workers in the MTurk sample also used CrowdFlower more than "a few times" and only 6.3% of workers in the CrowdFlower sample also used MTurk more than "a few times." The highest overlap was found in the sample of workers from Prolific Academic, where 22% reported also using MTurk.

Workers on other microtask platforms. A small number of studies have investigated the demographics of workers on other microtask platforms. For example, Hirth et al. (2011) examined the home countries of requesters and workers on the platform Microworkers and found that the platform's workforce was much more geographically diverse than MTurk's workforce. Bertschek et al. (2015) collected 408 responses from crowdworkers on two German microtask platforms. The crowdworkers in their sample were predominantly male, and compared to the German working population, they were younger and more likely to be single. Furthermore, they were highly educated and a majority of them were either in education or in employment. D. Martin et al. (2017) analyzed the socio-demographic characteristics of workers on the platforms Microworkers and Crowdee and compared them to a sample of MTurk workers. In their analysis, they grouped the locations of workers on the Microworkers platform into two groups: "Western countries," which included all workers from Europe, Oceania, and North America and "developing countries," which included all workers located in South America, Asia, and Africa. Their results indicated that the workers on Microworkers and Crowdee were younger than MTurk's workforce, predominantly male, and highly educated. Compared to workers in the "Western countries" group, workers in the "developing countries" group were younger and more educated, had a lower household income despite living in larger households, and spent more time on the platform. Berg et al. (2018) compared the socio-demographic characteristics of workers on five different platforms, including MTurk, Figure Eight, Clickworker, Prolific Academic, and Microworkers. The study differentiated between American and Indian workers on MTurk but did not conduct analyses at the country level for the other platforms.

In Section 3.3.1, this thesis extends existing work on the socio-demographic characteristics of the microtask workforce. It presents the results of a large survey of workers on Figure Eight, covering similar respondent numbers for ten diverse countries over two points in time. The analysis presented in this thesis constitutes the first country-level comparison of socio-demographic worker characteristics that goes beyond an analysis of American and Indian workers on MTurk, and it represents the most comprehensive scientific collection of socio-demographic worker characteristics on Figure Eight to date.

2.3.3 Motivations of the Workforce

Akin to research on the socio-demographic characteristics of the microtask workforce, most studies on the motivations of workers on microtask platforms have focused on the platform Amazon Mechanical Turk, and, consequently, on workers located in the U.S. and India. The results of these studies suggest that workers have different motivations for engaging in this type of work.

Ipeirotis (2010a) conducted an early study on the reasons that workers on MTurk had for participating on the platform. In a survey, workers were asked the multiple choice question *"Why do you complete tasks in Mechanical Turk?"* and offered six response options: *"Fruitful way to spend free time and get some cash (e.g., instead of watching TV)," "For 'primary' income purposes (e.g., gas, bills, groceries, credit cards)," "For 'secondary' income purposes, pocket change (for hobbies, gadgets, going out)," "To kill time," "I find the tasks to be fun,"* and *"I am currently unemployed, or have only a part time job."* As Kaufmann et al. (2011) noted, not all of these response options seem to correspond to a single motivational factor. The study found that most workers did not select the reason *"I find the tasks to be fun,"* or the reason *"To kill time."* The study further found that the responses differed between American and Indian workers: Notably, very few Indian workers selected the reason *"To kill time,"* more American workers than Indian workers selected the reason *"I find the tasks to be fun,"* and more Indian workers than American workers reported treating MTurk as their primary source of income.

Buhrmester et al. (2011) asked 187 workers on MTurk about the reasons they had for working on the platform and offered five reasons that workers ranked on a Likert-type scale: *"Enjoy doing interesting tasks," "To kill time," "To have fun," "To make money,"* and *"To gain self-knowledge."* In the study, *"Enjoy doing interesting tasks,"* was the survey item with the highest mean and the items *"To gain self-knowledge"* and *"To make money"* had the lowest means. Similarly, Litman et al. (2015) used these five survey items to measure the motivations of MTurk workers located in India ($N = 529$) and in the U.S. ($N = 207$). In contrast to the study conducted by Buhrmester et al. (2011), the study found that *"To make money,"* was the item with the highest mean in both samples and the item *"To kill time"* had the lowest mean in both samples.

The discrepancies of these findings might be, in part, explained by the small size of some of the samples, or, as Litman et al. (2015) suggested, the motivations of the workforce may have changed in the relatively short period of time between the studies. However, the discrepancies might also stem from the measurement instruments used for measuring the motivations. The studies relied on a single observed variable to measure each motivational dimension and did not evaluate the reliability or validity of their measurements, nor were they based on a theory of motivation.

Kaufmann et al. (2011) proposed a model for measuring crowdworker motivations that was based on different existing instruments, including Hackman and Oldham's Job Diagnostic Survey (Hackman and Oldham, 1980) and a model proposed by Lakhani and Wolf (2005) for measuring motivations of open source software developers. In their model, Kaufmann et al. (2011) differentiated between enjoyment based motivation, community based motivation, immediate payoffs, delayed payoffs, and social motivation. Using a sample composed of Indian and U.S. workers on MTurk ($N = 431$), the study found that the construct with the highest score was "immediate payoffs," i.e., payment, and that the constructs related to fun and enjoyment were ranked highly. The construct "pastime," defined as acting out of boredom or just to "kill time," correlated positively with household income and negatively with the weekly time spent on MTurk. Furthermore, the study found that workers who reported spending a lot of time on the platform may be motivated differently than workers who reported spending little time on the platform. Kaufmann et al. (2011) evaluated the internal consistency of the different subscales of the model, reporting Cronbach's alpha values between 0.74 and 0.94.

Hossain (2012) created a classification of motivations for participating on crowdsourcing platforms, listing potential extrinsic and intrinsic motivators and incentives. Antin and Shaw (2012) used a list experiment to analyze social desirability effects in self-reported motivations of workers on MTurk. Workers were offered four reasons for doing microtasks: *"to kill time," "to make extra money," "for fun,"* and *"because it gives me a sense of purpose."* In the experiment, workers were shown either all or only three of the reasons and asked to report how many of them they considered to be a motivation. The study found that workers located in the U.S. tended to over-report all four reasons while workers located in India tended to over-report the reasons *"sense of purpose"* and under-report *"to kill time"* and *"for fun."*

Brawley and Pury (2016) measured intrinsic motivation of American ($N = 225$) and Indian ($N = 132$) workers on MTurk. For measuring intrinsic motivation, they used an adapted version of two subscales of the Flow Dimension Scale (Webster et al., 1993) and reported a Cronbach's alpha of 0.74 for their adapted version of the scale. In their study, they found that intrinsic motivation was positively related to job satisfaction. Naderi et al. (2014) evaluated a 4-factor model for measuring extrinsic motivations of crowdworkers on a sample of American workers on MTurk ($N = 117$), using a subset of the items from the SDT-based *Work Extrinsic and Intrinsic Motivation Scale* (WEIMS) developed by Tremblay et al. (2009). In this 4-factor model, the items used in WEIMS to measure identified and integrated regulation were merged into a single factor, and the intrinsic motivation factor was omitted. The model was evaluated using CFA, and while the study reported a CFI in the acceptable range (> 0.90), RMSEA was high (> 0.08).

After a first version of the study presented in Section 3.3.2 was published (Posch, Bleier, and Strohmaier, 2017), Naderi (2018) adapted and extended the WEIMS-based 4-factor model of extrinsic motivations. The adapted scale was evaluated on three samples of workers on MTurk ($N = 170$, 90 and 86) and measures five motivational dimensions, three of which are measured by WEIMS items (amotivation, external regulation, and identified regulation).

W.-C. Chen et al. (2019) analyzed correlations between demographic characteristics, motivations, and participation of workers on four online labor platforms, including two platforms that focus on microtasks (MTurk ($N = 451$) and Microsoft's Universal Human Relevance System ($N = 1144$)). To measure motivations, workers were asked the questions *"What is the primary reason you do crowdsourcing?"* and *"What is the secondary reason you do crowdsourcing?"* Workers were offered five response options, from which they could select one for each question: *(1) "To earn money," (2) "To do something with my spare time," (3) "To be my own boss," (4) "To gain experience that could lead to future job opportunity,"* and *(5) "To learn new skills."* W.-C. Chen et al. presumed that these items measured three different motivational dimensions: monetary reward, self-determination, and self-improvement. Their results suggested that workers who had other options to earn income (e.g., due to higher education) were more likely to report a primary reason other than money, and that workers living in countries other than the U.S. were less likely to report money as their primary reason than workers located in the U.S. Furthermore, the results suggested that the workers' motivations differed across platforms.

Besides these quantitative studies on the motivations of the microtask workforce, there have also been a number of qualitative studies investigating the motivations of workers on MTurk. Gupta, Crabtree, et al. (2014) and Gupta, D. B. Martin, et al. (2014) conducted an ethnographic study of Indian workers on MTurk, analyzing their job satisfaction and the enjoyment they derived from working on tasks as well as various aspects of their working conditions such as education, infrastructure, and cost of living. D. B. Martin et al. (2014) conducted an ethnomethodological analysis of the content of Turker Nation, a forum for MTurk users. Their study found that users on Turker Nation saw their activity on MTurk primarily as work and considered payment to be an important factor.

Deng and Joshi (2016) asked 55 U.S.-based crowdworkers on MTurk a series of open-ended questions concerning different aspects of their work and analyzed the responses using revealed causal mapping. Based on concepts from Hackman and Oldham's job characteristics theory (Hackman and Oldham, 1975), they identified seven constructs that drive participation. These constructs included four motivational factors (crowdwork context, crowdsourcing task characteristics, crowdworker needs, and digital work control) as well as three socio-psychological outcomes (hedonic outcome, work value outcome, and crowdsourcing satisfaction outcome).

Jiang et al. (2015) conducted a survey with open-ended questions on MTurk, asking workers about the perceived benefits from working on the platform. Analyzing the workers' responses to this question, the study found that there were five categories of perceived benefits (monetary compensation, self-improvement, time management, emotional rewards, and task-characteristic benefits) and that Indian workers differed from American workers with respect to the perceived benefits. Furthermore, the results of the study suggested that workers compartmentalized the income from MTurk into different mental accounts.

A small number of studies have also attempted to manipulate workers' motivations via task framing and payment (Rogstadius et al., 2011; D. Chandler and Kapelner, 2013), achievement feedback (Lee et al., 2013), or by introducing time constraints and payments contingent on winning a contest (Feyisetan and Simperl, 2019).

In Section 3.3.2, this thesis extends existing work on the motivations of the microtask workforce by presenting the *Multidimensional Crowdworker Motivation Scale* (MCMS), a theory-based and cross-nationally applicable instrument for measuring the motivations of crowdworkers. Furthermore, Section 3.3.2 presents an analysis of crowdworker motivations in ten countries and three country income groups. In contrast to previous studies on the motivations of crowdworkers, we provide extensive evidence for the validity of our measurement instrument and we demonstrate cross-group comparability via measurement invariance tests (also see Section 2.1.2) prior to conducting any cross-group comparisons.

3 Publications

This chapter presents the publications contained in this cumulative thesis. First, Section 3.1 describes my contributions to the individual publications. Section 3.2 focuses on the use of the microtask workforce and presents three use cases, in each of which we employed microtasks in a different stage of the machine learning process. Section 3.3 focuses on the socio-demographic characteristics and motivations of the international microtask workforce.

3.1 Contributions to the Publications

This section describes my contributions to the individual publications contained in this cumulative thesis. In all publications, I was responsible for the design, implementation, and execution of all microtasks. In the following, I describe the details of my contributions to each publication.

- Posch, L., Bleier, A., Schaer, P., and Strohmaier, M. (2015). "The Polylingual Labeled Topic Model." In: *KI 2015: Advances in Artificial Intelligence*.

In the publication *"The Polylingual Labeled Topic Model,"* I was mainly responsible for the development of the conceptual framework as well as the development and implementation of the Polylingual Labeled Topic Model (PLL-TM), in collaboration with Arnim Bleier. I was further responsible for the evaluation of the model based on the semantic coherence of the topics and designed, implemented, and executed the microtasks necessary for this evaluation. The technical evaluation was conducted by me, in coordination with Arnim Bleier. Furthermore, I was responsible for the design and implementation of a visualization system based on the PLL-TM, which was published separately (Posch, Schaer, et al., 2016).

The idea for this publication stems from discussions between Arnim Bleier, Markus Strohmaier and me, and it was refined in discussions between Arnim Bleier, Philipp Schaer, Markus Strohmaier, and me. All authors contributed to the writing, reviewing, and editing of the manuscript.

- Stier, S., Posch, L., Bleier, A., and Strohmaier, M. (2017). "When populists become popular: Comparing Facebook use by the right-wing movement Pegida and German political parties." In: *Information, Communication & Society 20.9*.

In the publication *"When Populists Become Popular: Comparing Facebook Use by the Right-Wing Movement Pegida and German Political Parties,"* I was mainly responsible for the conceptualization of the technical framework for estimating the topics and their use by different parties over time, and I was responsible for the preprocessing of the data, the implementation of the framework, and the inference of the topics. Furthermore, I was responsible for designing, implementing, and executing the microtasks we used to calculate the topic salience weighted by populism over time. The labeling of the topics was conducted by Sebastian Stier, Arnim Bleier, and me.

The idea for this publication stems from discussions between Sebastian Stier, Arnim Bleier, Markus Strohmaier, and me. Sebastian Stier was responsible for the theoretical background and for the interpretation and analysis of the results in the context of political science. Furthermore, Sebastian Stier was responsible for retrieving the data from Facebook, for creating the visualizations, and for the calculation of the user overlap between the political parties. All authors contributed to the writing, reviewing, and editing of the manuscript.

- Eberhard, L., Walk, S., Posch, L., and Helic, D. (2019). "Evaluating narrative-driven movie recommendations on Reddit." In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*.

In the publication *"Evaluating Narrative-Driven Movie Recommendations on Reddit,"*, my main contribution was the design, implementation, and execution of the microtasks. We employed these microtasks to create a dataset for evaluating the potential of different recommender algorithms for calculating narrative-driven recommendations.

The original idea for this paper was developed by Lukas Eberhard, Simon Walk, and Denis Helic, and it was refined in discussions between Lukas Eberhard, Simon Walk, Denis Helic, and me. Lukas Eberhard, the main author of this publication, was primarily responsible for the design and implementation of the recommender framework, for the evaluation of the different recommender algorithms, and for the analysis of different post-filtering and re-ranking strategies. Simon Walk contributed to the design of the evaluation setup and to the compilation of the reference evaluation dataset. All authors contributed to the writing, reviewing, and editing of the manuscript.

- Posch, L., Bleier, A., Flöck, F., and Strohmaier, M. (2018). "Characterizing the global crowd workforce: A cross-country comparison of crowdworker demographics." *arXiv:1812.05948*.

In the publication *"Characterizing the Global Crowd Workforce: A Cross-Country Comparison of Crowdworker Demographics,"* I was responsible for the conceptualization and design of the study. I was further responsible for the design of the questionnaire and for the design, implementation, and execution of the microtasks. The analyses of the results were conducted by me, in coordination with Arnim Bleier and Fabian Flöck.

The idea for this publication stems from discussions between Arnim Bleier, Fabian Flöck, Markus Strohmaier, and me. Arnim Bleier contributed to the visualizations of the characteristics of the workforce.

All authors contributed to the writing, reviewing, and editing of the manuscript.

- Posch, L., Bleier, A., Lechner, C. M., Danner, D., Flöck, F., and Strohmaier, M. (2019). "Measuring motivations of crowdworkers: The Multidimensional Crowdworker Motivation Scale." In: *ACM Transactions on Social Computing* 2.2.

In the publication *"Measuring Motivations of Crowdworkers: The Multidimensional Crowdworker Motivation Scale,"* I was responsible for the conceptualization and implementation of the studies. I was responsible for evaluating existing models for work motivation in the context of microtasks, for the design of the item pool, and for the design of the final model. Furthermore, I was responsible for all data collections and for the design, implementation, and execution of all microtasks.

The reduction of the item pool and the validation of the factorial structure of the final model were conducted by me. The further evaluation of the different types of validity was conducted by me and conceptualized in discussions between Arnim Bleier, Clemens Lechner, Daniel Danner, and me. The evaluation regarding the applicability across platforms was conducted by me. The analysis of the cross-group comparability of the results was conducted by me, in close coordination with Arnim Bleier, Clemens Lechner, and Daniel Danner. The analysis of the results of the cross-country comparison was conducted by me, in collaboration with Arnim Bleier and Fabian Flöck.

The original idea for this publication stems from discussions between Arnim Bleier, Markus Strohmaier, and me, and it was refined in discussions between Arnim Bleier, Clemens Lechner, Daniel Danner, Fabian Flöck, Markus Strohmaier, and me. All authors contributed to the writing, reviewing, and editing of the manuscript.

3.2 Use of the Microtask Workforce

This section presents three use cases in which we employed the microtask workforce to support the analysis of large text corpora. In all presented use cases, automated methods alone did not suffice, and crowdworkers were indispensable for answering the respective research questions posed in the individual publications. Each of the presented use cases contributes to answering the first overarching research question of this thesis (*RQ1*, presented in Section 1.4) by demonstrating ways in which human input from microtasks complements automated methods for the analysis of large text corpora in different stages of the machine learning process.

In each of the use cases, microtasks were employed in a different stage of the machine learning process. In the publication presented in Section 3.2.1, we developed a new topic model for multilingual, labeled documents. In addition to a technical evaluation of the model, an evaluation of the semantic coherence of the topics was necessary to determine the quality of the model compared to existing models. As no automated method to reliably determine semantic coherence exists, we evaluated the model with the help of human input from crowdworkers. Section 3.2.2 presents a study on populist political communication in online social media. The study examines the topics addressed by different German political parties and by the movement Pegida. As no reliable automated method exists for measuring populism in text, we employed the microtask workforce to interpret the model's parameters in the context of populist communication. In the publication presented in Section 3.2.3, we evaluated the utility of commonly used recommender algorithms with respect to their potential for incorporating information from narrative descriptions of users' preferences into the recommendations. We employed the microtask workforce during the preprocessing of the unstructured text data containing narrative descriptions of preferences, for the task of extracting structured data such as named entities, relevant contextual information, and user sentiment.

3.2.1 The Polylingual Labeled Topic Model

This article presents the *Polylingual Labeled Topic Model (PLL-TM)*, a new topic model that combines the characteristics of two existing topic models, *Labeled LDA* and the *Polylingual Labeled Topic Model*. We developed the PLL-TM for measuring latent topics in corpora consisting of unstructured text documents that are present in multiple languages and that are labeled according to a classification system.

In the article, we present the model's generative storyline as well as an inference strategy based on Gibbs sampling to estimate the topics. We applied the PLL-TM to a corpus consisting of documents from the social science domain, in a two-language setting: The natural language German represented the first language, and the controlled vocabulary of the *Thesaurus for the Social Sciences* (Zapilko et al., 2013) represented the second language.

We employed microtasks in the *model evaluation* stage of the machine learning process. Specifically, we compared the proposed PLL-TM's performance on the corpus to that of three existing topic models: LDA, L-LDA, and the PLTM. With the help of microtasks, we evaluated the semantic coherence of the topics estimated by the different topic models. In addition, we performed a technical evaluation of the different models via perplexity. The results of the evaluation showed that the PLL-TM achieved not only a good predictive performance but also produced topics that had a high semantic coherence.

Based on the PLL-TM presented in this article, we developed a visualization system for creating and visualizing probabilistic semantic links between thesaurus descriptors and classes contained in a classification system, which was published separately (Posch, Schaer, et al., 2016).

The Polylingual Labeled Topic Model

Lisa Posch[1,2]( ), Arnim Bleier[1], Philipp Schaer[1], and Markus Strohmaier[1,2]

[1] GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany
{lisa.posch,arnim.bleier,philipp.schaer,markus.strohmaier}@gesis.org
[2] Institute for Web Science and Technologies,
University of Koblenz-Landau, Mainz, Germany

Abstract. In this paper, we present the *Polylingual Labeled Topic Model*, a model which combines the characteristics of the existing *Polylingual Topic Model* and *Labeled LDA*. The model accounts for multiple languages with separate topic distributions for each language while restricting the permitted topics of a document to a set of predefined labels. We explore the properties of the model in a two-language setting on a dataset from the social science domain. Our experiments show that our model outperforms LDA and Labeled LDA in terms of their held-out perplexity and that it produces semantically coherent topics which are well interpretable by human subjects.

Keywords: Thesauri · Classification · Probabilistic linking · Topic models

1 Introduction

Topic models are a popular and widely used method for the analysis of textual corpora. *Latent Dirichlet Allocation (LDA)* [2], one of the most popular topic models, has been adapted to a multitude of different problem settings, such as modeling labeled documents with *Labeled LDA (L-LDA)* [9] or modeling multilingual documents with *Polylingual Topic Models (PLTM)* [7]. Textual corpora often exhibit both of these characteristics, containing documents in multiple languages which are also annotated with a classification system. However, there is currently no topic model which possesses the ability to process multiple languages while simultaneously incorporating the documents' labels.

To close this gap, this paper introduces the *Polylingual Labeled Topic Model (PLL-TM)*, a model which combines the characteristics of PLTM and L-LDA. PLL-TM models multilingual labeled documents by generating separate distributions over the vocabulary of each language, while restricting the permitted topics of a document to a set of predefined labels. We explore the characteristics of our model in a two-language setting, with German natural language text as the first language and the controlled *SKOS* vocabulary of a thesaurus as the second language. The labels of the documents, in our setting, are classes from the classification system with which our corpus is annotated.

S. Hölldobler et al. (Eds.): KI 2015, LNAI 9324, pp. 295–301, 2015.
DOI: 10.1007/978-3-319-24489-1_26

Contributions. The main contribution of this paper is the presentation of the PLL-TM. We present the model's generative storyline as well as an easy-to-implement inference strategy based on Gibbs sampling. For evaluation, we compute the held-out perplexity and conduct a word intrusion task with human subjects using a dataset from the social science domain. On this dataset, the PLL-TM outperforms LDA and L-LDA in terms of its predictive performance and generates semantically coherent topics. To the best of our knowledge, PLL-TM is the first model which accounts for multiple vocabularies and, at the same time, possesses the ability to restrict the topics of a document to its labels.

2 Related Work

Topic models are generative probabilistic models for discovering latent topics in documents and other discrete data. One of the most popular topic models, LDA, is a generative Bayesian model which was introduced by Blei et al. [2]. In this section, we review LDA, as well as the two other topic models whose characteristics we are going to integrate into PLL-TM.

LDA. Beginning with LDA [2], we follow the common notation of a document d being a vector of N_d words, $\boldsymbol{w}_d$, where each word w_{di} is chosen from a vocabulary of V terms. A collection of documents is defined by $\mathcal{D} = \{\boldsymbol{w}_1,...,\boldsymbol{w}_D\}$. LDA's generative storyline can be described by the following steps.

1. For each document $d \in \{1,...,D\}$, a distribution θ_d over topics is drawn from a symmetric K-dimensional Dirichlet prior parametrized by α:

$$\theta_d \sim Dir(\alpha) \ . \tag{1}$$

2. Then, for each topic $k = \{1,...,K\}$, a distribution ϕ_k over the vocabulary is drawn form a V-dimensional Dirichlet distribution parametrized by β:

$$\phi_k \sim Dir(\beta) \ . \tag{2}$$

3. In the final step, the i^{th} word in document d is generated by first drawing a topic index z_{di} and subsequently, a word w_{di} from the topic indexed by z_{di}:

$$w_{di} \sim Cat(\phi_{z_{di}}) \ , \qquad z_{di} \sim Cat(\theta_d) \ . \tag{3}$$

Labeled LDA. Ramage et al. [9] introduced L-LDA, a supervised version of LDA. In L-LDA, a document d's topic distribution θ_d is restricted to a subset of all possible topics $\boldsymbol{\Lambda}_d \subseteq \{1,..,K\}$. Here, collection of documents is defined by $\mathcal{D} = \{(\boldsymbol{w}_1,\boldsymbol{\Lambda}_1),...,(\boldsymbol{w}_D,\boldsymbol{\Lambda}_D)\}$. The first step in L-LDA's generative storyline draws the distribution of topics θ_d for each document $d \in \{1,...,D\}$

$$\theta_d \sim Dir(\alpha\boldsymbol{\mu}_d) \ , \tag{4}$$

where α is a continuous positive valued scalar and $\boldsymbol{\mu}_d$ is a K-dimensional vector

$$\mu_{dk} = \begin{cases} 1 & \text{if } k \in \boldsymbol{\Lambda}_d \\ 0 & \text{otherwise} \ , \end{cases} \tag{5}$$

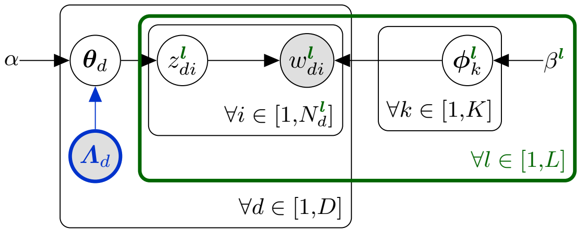


Fig. 1. The PLL-TM in plate notation. Random variables are represented by nodes. Shaded nodes denote the observed words and labels, bare symbols indicate the fixed priors α and β^l. Directed edges between the nodes then define conditional probabilities, where the child node is conditioned on its parents. The rectangular plates indicate replication over data-points and parameters. Colors indicate the parts which are inherited from L-LDA (**blue**) and PLTM (**green**). Black is used for the LDA base.

indicating which topics are permitted. Once these label-restricted topic distributions are drawn, the process of generating documents continues identically to the generative process of LDA. In the case of $\boldsymbol{\Lambda}_d = \{1,..,K\}$ for all documents, no restrictions are active and L-LDA is reduced to LDA.

Polylingual Topic Model. Ni et al. [8] extended the generative view of LDA to multilingual documents. Mimno et al. [7] elaborated on this concept, introducing the *Polylingual Topic Model* (PLTM). PLTM assumes that the documents are available in L languages. A document d is represented by $[\boldsymbol{w}^1_d,...,\boldsymbol{w}^L_d]$, where for each language $l \in 1,...,L$, the vector $\boldsymbol{w}^l_d$ consists of N^l_d words which are chosen from a language specific vocabulary with V^l terms. A collection of documents is then defined by $\mathcal{D} = \{[\boldsymbol{w}^1_1,...,\boldsymbol{w}^L_1],...,[\boldsymbol{w}^1_D,...,\boldsymbol{w}^L_D]\}$. The generative storyline is equivalent to LDA's except that steps 2 and 3 are repeated for each language. Hence, for each topic $k = \{1,...,K\}$ in each language $l \in \{1,...,L\}$, a language specific topic distribution ϕ^l_k over the vocabulary of length V^l is drawn:

$$\phi^l_k \sim Dir(\beta^l) \ . \tag{6}$$

Then, the i^{th} word of language l in document d is generated by drawing a topic index z^l_{di} and subsequently, a word w^l_{di} from a language specific topic distribution indexed by z^l_{di}:

$$w^l_{di} \sim Cat(\phi^l_{z^l_{di}}) \ , \qquad\qquad z^l_{di} \sim Cat(\theta_d) \ . \tag{7}$$

Note that in the special case of just one language, i.e. $L = 1$, PLTM is reduced to LDA.

3 The Polylingual Labeled Topic Model

In this section, we introduce the *Polylingual Labeled Topic Model (PLL-TM)*, which integrates the characteristics of the models described in the previous section into a single model. Figure 1 depicts the PLL-TM in plate notation. Here, a collection of documents is defined by $\mathcal{D} = \{[\boldsymbol{w}_1^1,...,\boldsymbol{w}_1^L],\boldsymbol{\Lambda}_1)),...,[\boldsymbol{w}_D^1,...,\boldsymbol{w}_D^L],\boldsymbol{\Lambda}_D)\}$.

The generative process follows three main steps:

1. For each document $d \in \{1,...,D\}$, we draw the distribution of topics

$$\theta_d \sim Dir(\alpha \boldsymbol{\mu}_d) \ , \tag{8}$$

where $\boldsymbol{\mu}_d$ is computed according to Equation 5.

2. For each topic $k \in \{1,...,K\}$ in each language $l \in \{1,...,L\}$, we draw a distribution over the vocabulary of size V^l:

$$\phi_k^l \sim Dir(\beta^l) \ , \tag{9}$$

3. Next, for each word in each language l of document d, we draw a topic

$$w_{di}^l \sim Cat(\phi^l_{z^l_{di}}) \ , \qquad\qquad z_{di}^l \sim Cat(\theta_d) \ . \tag{10}$$

Note that PLL-TM contains both PLTM and L-LDA as special cases.

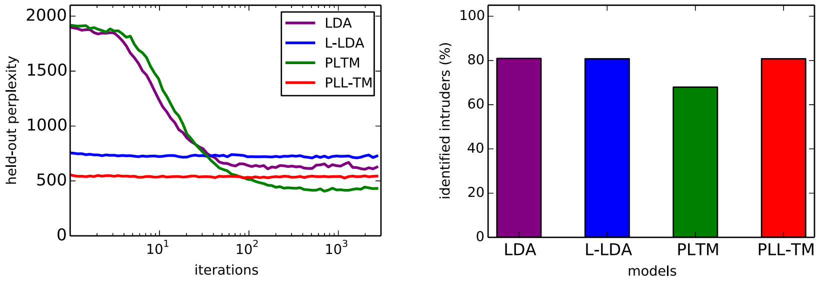
For inference, we use collapsed Gibbs sampling [6] for the indicator variables $\boldsymbol{z}$, with all other variables integrated out. The full conditional probability for a topic k is given by

$$P(z_{di}^l = k \mid w_{di}^l = t,...) \propto \frac{n_{dk}^{\neg di} + \alpha}{n_{d.}^{\neg di} + K\alpha} \times \frac{n_{kt}^{l\neg di} + \beta^l}{n_{k.}^{l\neg di} + V^l \beta^l} \ , \tag{11}$$

where n_{dk} is the number of tokens allocated to topic k in document d, and n_{kt}^l is the number of tokens of word $w_{di}^l = t$ which are assigned to topic k in language l. Furthermore, $\cdot$ is used in place of a variable to indicate that the sum over its values (i.e. $n_{d.} = \sum_k n_{dk}$) is taken and $\neg di$ to mark the current token as excluded. While the full conditional posterior distribution is reminiscent of the one used in PLTM, the assumptions of the L-LDA model restrict the probability $P(z_{di}^l = k)$ to those $k \in \boldsymbol{\Lambda}_d$ with which document d is labeled.

Table 1. This table shows the five most probable terms for two classes in the CSS, generated by PLL-TM, in two languages: *TheSoz* (TS) and German natural language words with their translation (AB).

	Population Studies, Sociology of Population:
TS:	*population development*, *demographic aging*, *population*, *demographic factors*, *demography*
AB:	wandel, demografischen, bevlkerung, deutschland, entwicklung (change, demographic, population, germany, development)
	Developmental Psychology:
TS:	*child*, *developmental psychology*, *adolescent*, *personality development*, *socialization research*
AB:	entwicklung, sozialisation, kinder, kindern, identitt (development, socialization, children, children, identity)



(a) Comparison of the held-out perplexity (lower values are better) as a function of iterations.

(b) Comparison of the semantic coherence (word intrusion) of the generated topics.

Fig. 2. Evaluation of the PLL-TM. These figures show that on the SOLIS dataset, PLL-TM outperforms LDA and L-LDA in terms of its predictive performance and produces topics with a higher semantic coherence than PLTM.

4 Evaluation

For our evaluation, we use documents from the *Social Science Literature Information System (SOLIS)*. The documents are manually indexed with the SKOS *Thesaurus for the Social Sciences (TheSoz)* [10] and manually classified with the *Classification for the Social Sciences (CSS)* by human domain experts. For our experiments, we used all SOLIS documents which were published in the years 2008 to 2013, resulting in a corpus of about 60.000 documents.

We explore the characteristics of our model in a two-language setting, with German natural language text as the first language ($AbstractWords$) and the controlled *SKOS* vocabulary of a thesaurus as the second language ($TheSoz$). The labels of the documents, in our setting, are classes from the CSS. After applying standard preprocessing to remove rare words and stopwords, $TheSoz$ consisted of 802.764 tokens over a vocabulary of 7.406 distinct terms, and $AbstractWords$ consisted of 5.417.779 tokens over a vocabulary of about 43.000 distinct terms. In our corpus, each document is labeled with an average of 2.14 classes.

We compare four different topic models: LDA, L-LDA, PLTM and PLL-TM. The unilingual models (i.e. LDA and L-LDA) were trained on language $TheSoz$; the polylingual models (i.e. PLTM and PLL-TM) were trained on $TheSoz$ and $AbstractWords$. The documents in our corpus were labeled with a total of 131 different classes from the CSS and we trained the unlabeled models with an equal number of topics. α and β^l were specified with 0.1 and 0.01, respectively. Table 1 shows the topics generated by PLL-TM for two classes of the CSS, reporting the five most probable terms for the languages $TheSoz$ and $AbstractWords$.

Language Model Evaluation. For an evaluation of the predictive performance, we computed the held-out perplexity for all models. We held out 1.000

documents as test set $\mathcal{D}_{test}$ and, with the remaining data $\mathcal{D}_{train}$, we trained the four models. We split each test document in the following way:

- $\boldsymbol{x}_{d1}$: All words of language $AbstractWords$ and a randomly selected 50% of the words in language $TheSoz$ which occur in document d.
- $\boldsymbol{x}_{d2}$: The remaining 50% of the words in language $TheSoz$ which occur in document d.

The test documents for the unilingual models were split analogously, with $\boldsymbol{x}_{d1}$ consisting of 50% percent of the words in language $TheSoz$ which occur in document d. For each document d, we computed the perplexity of $\boldsymbol{x}_{d2}$.

Figure 2a shows the results of this evaluation. One can see that the labeled models both start out with a lower perplexity and need less iterations to achieve a good performance, which is due to the fact that the labels provide additional information to the model. In contrast, the unlabeled models need almost 100 iterations to achieve a comparable performance. On our corpus, PLL-TM outperformed LDA and L-LDA, and even though PLL-TM had a higher perplexity than PLTM, it is important to keep in mind that PLTM does not possess the ability to produce topics which correspond to the classes of the CSS.

Human Evaluation of the Topics. Chang et al. [4] proposed a formal setting in which humans evaluate the latent space of a topic model. For evaluating the topics' semantic coherence, they proposed a *word intrusion* task: Crowdworkers were shown six terms, five of which were highly probable terms in a topic and one was an "intruder" – an improbable term for this topic which had a high probability in some other topic.

We conducted the word intrusion task for the four topic models on CrowdFlower [1], with ten distinct workers for each topic in each model. Figure 2b shows the results of this evaluation for the different models. For each model, the figure depicts the percentage of topics for which the ten workers collectively detected the correct intruder. The collective decision was based on CrowdFlower's *confidence score*, i.e. the level of agreement between workers weighted by each worker's percentage of correctly answered test questions. The results show that PLL-TM produces topics which are equally coherent as unilingual models, and more coherent than the topics produced by PLTM.

5 Discussion and Conclusions

In this paper, we presented PLL-TM, a joint model for multilingual labeled documents. The results of our evaluation showed that PLL-TM was the only model which produced both highly interpretable topics and achieved a good predictive performance. Compared to L-LDA, the only other model capable of incorporating label information, our model produced equally well interpretable topics while achieving a better predictive performance. Compared to PLTM, the only other model capable of dealing with multiple languages, PLL-TM had a lower predictive performance, but produced topics with a higher semantic coherence. For future work, we plan an evaluation of the model in a label prediction task and an application of the model in a setting with more than two natural languages. Furthermore, we plan an evaluation on a larger dataset using a more memory-friendly inference strategy such as *Stochastic Collapsed Variational Bayesian Inference* [5], which has been shown to be applicable outside of its original LDA application [3].

References

1. Biewald, L.: Massive multiplayer human computation for fun, money, and survival. In: Harth, A., Koch, N. (eds.) ICWE 2011. LNCS, vol. 7059, pp. 171–176. Springer, Heidelberg (2012)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3**, 993–1022 (2003)
3. Bleier, A.: Practical collapsed stochastic variational inference for the hdp. In: NIPS Workshop on Topic Models: Computation, Application, and Evaluation (2013)
4. Chang, J., Boyd-Graber, J.L., Gerrish, S., Wang, C., Blei, D.M.: Reading tea leaves: how humans interpret topic models. In: Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held December 7–10, 2009, Vancouver, British Columbia, Canada, pp. 288–296 (2009)
5. Foulds, J.R., Boyles, L., DuBois, C., Smyth, P., Welling, M.: Stochastic collapsed variational bayesian inference for latent dirichlet allocation. In: The 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2013, Chicago, IL, USA, pp. 446–454, August 11–14, 2013
6. Griffiths, T.L., Steyvers, M.: Finding scientific topics. In: Proceedings of the National Academy of Sciences (2004)
7. Mimno, D.M., Wallach, H.M., Naradowsky, J., Smith, D.A., McCallum, A.: Polylingual topic models. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, August 6–7, 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 880–889 (2009)
8. Ni, X., Sun, J., Hu, J., Chen, Z.: Mining multilingual topics from wikipedia. In: Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, pp. 1155–1156, April 20–24, 2009
9. Ramage, D., Hall, D.L.W., Nallapati, R., Manning, C.D.: Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, August 6–7, 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 248–256 (2009)
10. Zapilko, B., Schaible, J., Mayr, P., Mathiak, B.: Thesoz: A SKOS representation of the thesaurus for the social sciences. Semantic Web **4**(3), 257–263 (2013)

3.2.2 When Populists Become Popular: Comparing Facebook Use by the Right-Wing Movement Pegida and German Political Parties

In this article, we present an analysis of online political communication by German political parties and by the right-wing populist movement Pegida. For the analysis, we modeled the topics addressed by different political actors on the social media platform Facebook over a period of around 20 months, starting on the day of the opening of the Pegida Dresden account on 29 December 2014. Specifically, we used LDA to estimate party-specific topic probabilities and then analyzed the extent to which the different political groups discussed different topics as well as how these topic mixtures changed over time.

For an interpretation of the model in the context of populist communication, we employed microtasks in the *model interpretation* stage of the machine learning process. In these microtasks, crowdworkers interpreted the model parameters, i.e., the estimated topics, and judged the extent to which they corresponded to the three criteria of populist communication defined by Reinemann et al. (2016).

The results of the analysis showed that the movement Pegida and the party AfD emphasized populist topics more than other parties, while the governing parties CDU and SPD tended to de-emphasize those topics. Other opposition parties engaged in populist communication to varying degrees. Furthermore, Pegida and AfD had the highest similarity in topic distributions of all pairs of political groups, except for the sister parties CDU and CSU, which had equally similar topic distributions.

INFORMATION, COMMUNICATION & SOCIETY, 2017
VOL. 20, NO. 9, 1365–1388
https://doi.org/10.1080/1369118X.2017.1328519

Routledge
Taylor & Francis Group

When populists become popular: comparing Facebook use by the right-wing movement Pegida and German political parties

Sebastian Stier[a], Lisa Posch[a,b], Arnim Bleier[a] and Markus Strohmaier[a,b]

[a]Department Computational Social Science, GESIS – Leibniz Institute for the Social Sciences, Cologne, Germany; [b]Department of Computer Science, University of Koblenz-Landau, Koblenz, Germany

ABSTRACT

Previous research has acknowledged the use of social media in political communication by right-wing populist parties and politicians. Less is known, however, about its pivotal role for right-wing social movements which rely on personalized messages to mobilize supporters and challenge the mainstream party system. This paper analyzes online political communication by the right-wing populist movement Pegida and German political parties. We investigate to which extent parties attract supporters of Pegida, to which extent they address topics similar to Pegida and whether their topic use has become more similar over a period of almost two years. The empirical analysis is based on Facebook posts by main accounts and individual representatives of these political groups. We first show that there are considerable overlaps in the audiences of Pegida and the new challenger in the party system, AfD. Then we use topic models to characterize topic use by party and surveyed crowdworkers to which extent they perceive the identified topics as populist communication. The results show that while Pegida and AfD talk about rather unique topics and smaller parties engage to varying degrees with the topics populists emphasize, the two governing parties CDU and SPD clearly deemphasize those. Overall, the findings indicate that the considerable attention devoted to populist actors and shifts in public opinion due to the refugee crisis have left only moderate marks in political communication within the mainstream party system.

ARTICLE HISTORY
Received 29 November 2016
Accepted 4 May 2017

KEYWORDS
Political communication; populism; protest movements; party politics; topic modeling; Facebook

Introduction

Right-wing populist forces are challenging the established political order across the Western world. Previous research on populism has acknowledged the pivotal role of social media in these processes, enabling populist parties and politicians to bypass media gatekeepers and transmit direct messages to target audiences (Arzheimer, 2015; Engesser, Ernst, Esser, & Büchel, 2016). At the same time, the technological opportunity structures of right-wing social movements like the Tea Party or the alt-right in the U.S. have also improved significantly since the advent of social media. These processes are less well understood, since previous social movement research mostly focused on how leftist and

CONTACT Sebastian Stier sebastian.stier@gesis.org
Supplemental data for this article can be accessed at https://doi.org/10.1080/1369118X.2017.1328519

anti-authoritarian groups communicate, mobilize and organize collective action (e.g. Bennett, Segerberg, & Walker, 2014; González-Bailón & Wang, 2016). Furthermore, the socio-technical characteristics of social media make it a unique venue for direct interactions between social movements, their supporters and parties, a core mechanism how emerging societal ideas are established in democracies (McAdam & Tarrow, 2010). To improve the understanding of these recent phenomena, we concentrate on contemporary German politics.

In October 2014, the protest movement Pegida (Patriotic Europeans against the islamization of the West) emerged in Dresden demonstrating, *inter alia*, against 'islamization', 'unchecked mass immigration', 'genderization', international trade treaties and further European integration of nation states (Pegida, 2015). Offshoots of the movement formed in cities across the country and Pegida immediately received a great share of media attention. But Pegida was not the only political newcomer in German politics: The AfD (Alternative for Germany) was founded in February 2013, has since shifted to the right and achieved striking electoral successes in elections for the European and state parliaments. First systematic analyses of political scientists classified the AfD as 'right-wing populist' (Arzheimer, 2015; Berbuir, Lewandowsky, & Siri, 2015) although the party defines itself as a party in the center of the ideological spectrum, a self-described *Volkspartei*.

In light of this populist surge and an anxious public due to the refugee crisis (infratest dimap, 2016), the established political parties faced considerable pressures to adapt. The CSU, for instance, has urged chancellor Angela Merkel and the CDU to make their liberal refugee policies much more restrictive. If not in policies, parties could at least adjust their political communication by picking up topics emphasized by populists in order to demonstrate responsiveness. In the German case, Facebook is particularly well suited for such personalized messages to citizens since it is the social network with the highest societal diffusion (Frees & Koch, 2015). Considering the interactive nature of Facebook, especially politicians who frequently encounter users with populist leanings on their pages might be inclined to adjust their messages. In order to investigate these dynamics, we pose the following research questions.

1 To which extent do the audiences of Pegida and German political parties overlap on Facebook?
2 To which extent do Pegida and German political parties discuss similar topics?
3 Does communication by German political parties increasingly converge to topics emphasized by populists?

Theoretically, the paper discusses the related literatures on online collective action, populism and party competition in Western democracies. Empirically, we analyze Facebook posts by party accounts and individual representatives of Pegida and German political parties over a period of almost two years, starting in December 2014. First, we concentrate on overlaps in audiences according to two Facebook conventions: likes and comments. We then train topic models on all posts to estimate party specific topic probabilities and survey crowdworkers to which extent they perceive the identified topics as populist communication. In the analysis, we find considerable similarities between Pegida and AfD in terms of audiences and topics in comparison to other German parties. Smaller parties engage to varying degrees with the topics populists emphasize, however, the governing parties CDU and SPD clearly deemphasize those. Overall, the findings indicate that the considerable attention devoted to populist actors and shifts in public opinion due to the refugee crisis have left only moderate marks in political communication within the mainstream party system.

Pegida, AfD and populist tendencies in Germany

In this section, we provide background information on the recent emergence of populist groups in Germany and shifts in public opinion. We first focus on Pegida which started as a local protest movement meeting for demonstrations each Monday in Dresden. It emerged from a non-public Facebook discussion group created on 11 October 2014. The movement reached its height in participation in January 2015 when up to 25,000 people participated in one Monday demonstration and branches had been founded in most major German cities (Vorländer, Herold, & Schäller, 2016). Since then, the size of the crowds at Pegida's demonstrations and public attention to the movement have varied considerably. Yet, Pegida regularly enters public debates, recently in October 2016, when its activists protested against the assembled political and media elites of the country who came to Dresden to celebrate the German national holiday.

Facebook is the main platform for Pegida to present its political opinions and organize collective action (Vorländer et al., 2016). As Pegida mostly refuses to talk to traditional media, its Facebook pages are the most exhaustive and, besides the speeches at the Monday demonstrations, the only textual manifestation of its policy positions. In contrast, Pegida did not create a Twitter account until January 2016. Arzheimer (2015) reports a similar preference for Facebook in case of the AfD. He relates this to the higher degree of control that an owner of a Facebook page can exert while debates on Twitter are publicly open and not subject to moderation (Arzheimer, 2015, p. 548). In terms of demographics, Facebook is a medium used by a considerable share of the German population, on a daily basis by 22% of Internet users while Twitter use in Germany still remains a 'special interest' (Frees & Koch, 2015). This makes Facebook a more attractive medium for populist online communication. Accordingly, Pegida became increasingly active on Facebook and still attracts significant numbers of supporters online (Vorländer et al., 2016, pp. 21–22; see section 'Results and discussion').

The AfD was founded by economics professors, businessmen and former members of the conservative and liberal parties CDU and FDP in 2013. It had predominantly been a Eurosceptic party criticizing the fiscal and monetary policies of the German government and the EU institutions during the Euro crisis (Arzheimer, 2015). But the AfD increasingly incorporated ideas critical of migration and Islam into its platform. This transition was accompanied by disputes between a nationalist and a liberal party wing, which ultimately led to a mass exodus of the latter, including party founder Bernd Lucke who proceeded to found the party ALFA. The internal disagreement about the official party position towards Pegida was among the central reasons for the party split (Korsch, 2016; Vorländer et al., 2016, pp. 39–43). These internal and public disputes notwithstanding, the AfD had considerable successes in the elections for the European parliament in 2014 gaining 7.1% of the German votes and winning seats in most elections for parliaments in the federal states since 2013. These electoral successes further strengthened the conservative forces in the party.

These two groups are part of a more general surge in populist sentiment in the country, expressed through polarized discussions on social media, high poll numbers for the AfD (up to 15% in national polls) and the party's electoral successes. Ongoing media coverage of protest events and their comments on the refugee crisis kept populist actors and positions on the public agenda. Furthermore, the refugee crisis laid bare considerable anti-immigration preferences in public opinion (infratest dimap, 2016). A recent study of Pegida makes the argument that the movement merely mirrors (preexisting) preferences of many citizens disconnected from the political system and its elites (Hein, 2017). In focusing on Pegida's social media activity, we thus not only aim to capture the communication of the movement, but regard it as a proxy for the political opinions of considerable segments of the German population.

Taken together, these developments have incited intense discussions in all German parties on how to address the grievances of citizens sympathizing with populist positions and actors.

Related research and expectations

The article is related to three rich research fields: social media use by protest movements, populist political communication and the literature on party competition. We review findings relevant for the present research questions and discuss expectations for our case.

Social media use by social movements and populists

Numerous studies focus on 'connective action' (Bennett et al., 2014), i.e. the use of social media for the mobilization of (loose) social movements (e.g. Bennett et al., 2014; González-Bailón & Wang, 2016). This field mostly concentrates on successful mobilization periods like the Arab Spring or the Occupy Wall Street and 15M movements. However, protest movements also have more latent and potentially long-lasting effects on political systems after news cycles have moved on. Since 'elections and social movements are the two major forms of political conflict in democratic systems' (McAdam & Tarrow, 2010, p. 532), the latter have historically influenced policies and party systems (McAdam & Tarrow, 2010). For instance, while the Occupy Wall Street movement has waned after a few months without leading to abrupt political changes, its message on economic inequality nonetheless still influences the debate on economic policies (Bennett, 2012) and its activists contributed to Bernie Sanders' insurgent campaign in the Democratic presidential primaries (CNN, 2016).

Since social movements often articulate grievances resonating with significant shares of the population, they can have an impact even if they are unable to maintain a prominent public profile. An increasing salience of their core issues and shifts in public opinion create electoral incentives for political parties to adjust their party programs accordingly (Downs, 1957). In the interactive communication environment of social media, political communication of social movements, parties and ordinary citizens is meshed together much more fluidly than in a mass media setting with traditional gatekeepers (Chadwick, 2013). The direct exposure of party actors to genuine political communication from the grassroots might accelerate contagion mechanisms identified by McAdam and Tarrow (2010) such as the penetration and lobbying of parties by social movements (*proactive movement* *mobilization*) and the emulation of interactive communication strategies by political parties themselves (*transferable innovations*).

There are indications that social media is especially beneficial to movements and parties on the political right, a tendency that has not yet been picked up by the connective action literature. Online social networks allow populists to bypass traditional media gatekeepers and use more radical rhetoric than previously possible in the age of mass media (Engesser et al., 2016). Right-wing populist communication appeals directly to the people and is particularly suited for personalized frames tailored to the interactive user experience: '[…] these late modern hybrids invite followers to define "true citizens" as "people like me" (e.g. a white, hard-working native-born citizen) and not those immigrants who come to live off my hard-earned tax money' (Bennett, 2012, p. 23).

The narrow focus on social media use by movements from the political left needs to be reconsidered and synchronized with established research on populism (Bale, Green-Pedersen, Krouwel, Luther, & Sitter, 2010; Mudde, 2004; Reinemann, Aalberg, Esser, Strömbäck & de Vreese, 2016). In order to do this, we rely on a minimalist conceptualization of populism. According to Reinemann et al. (2016), populism emphasizes (1) the political will of the people, (2) criticizes political or economic elites and (3) agitates against 'out groups' like religious or ethnic minorities. As a 'thin ideology' (Mudde, 2004) populism is compatible with diverse ideologies and is used as a communication mode by political actors from across the political spectrum (Mudde, 2004; Reinemann et al., 2016).

Within the German party system, the AfD still characterizes itself publicly as a party from the middle of the ideological spectrum. Nonetheless, media commentators described Pegida and like-minded citizens as the new target groups of the AfD after the Euro crisis lost its momentum as the main mobilizing issue (Die Welt, 2015). And the assessment of political scientists is also clear: 'Up to now the party in its ambiguity with links to both the self-declared "centre of society" and the far right is a functional equivalent for right-wing populism in Germany' (Berbuir et al., 2015, p. 174). Korsch (2016) recounts the evolving relationship between Pegida and AfD as one mostly determined by the internal struggles between competing factions within the AfD. He shows that since 2015, the increasingly dominant conservative party wing around Alexander Gauland and Björn Höcke advocated for a rapprochement between both groups. This tendency has accelerated in 2016 when Pegida openly advertised AfD contents on its Facebook page. We thus (1) expect that the AfD attracts the highest share of Pegida supporters and (2) that the AfD emphasizes populist issues in order to attract the voters with preferences similar to Pegida activists.

Party competition and populism

The literature on party competition in Western democracies (that is, however, negligent of social movements) can help us to elaborate on the question if established parties engage with or avoid topics populists typically emphasize. Of particular importance here is the debate on issue competition versus positional competition (Green-Pedersen, 2007). According to models of issue competition, also called saliency theory (Budge & Farlie, 1983), parties stress the topics that fall within their core competency while avoiding issues on which their competitors are seen as more competent (see also 'issue ownership' theory). On the other hand, the positional competition model states that parties compete against each other emphasizing similar issues while proposing different solutions (Dolezal, Ennser-Jedenastik, Müller, & Winkler, 2014; Downs, 1957). The core question in the context of our research question is therefore, whether parties enter or avoid issue areas introduced by populist challengers (Meguid, 2005; Mudde, 2004).

Meguid (2005) proposes a model of electoral success positing that established parties should either ignore topics stressed by niche parties in order to reduce their public salience or occupy them with converging positions. According to this logic, for which the author finds support using data from the Comparative Manifesto Project (CMP), niche parties, here populists, only gain electoral support if mainstream parties address their issues but with diverging positions. Additional studies confirmed that especially parties from the moderate right tend to adopt topics and positions by emerging right-wing (populist) parties, e.g. on immigration (Abou-Chadi, 2016; Bale, 2003). Other empirical studies, however, revealed an even more complex picture dependent upon the specifics of each party system. Bale et al. (2010) found that while social democratic parties in four European countries reacted to right-wing challengers with programmatic adjustments, these were far from uniform and confined by country-specific factors. Meanwhile, Rooduijn, de Lange, and van der Brug (2014) could not identify shifts in party positions in five Western European countries as a reaction to populist challenges. In terms of reactions by parties to shifts in public opinion, Adams, Clark, Ezrow, and Glasgow (2006) showed that in contrast to mainstream parties, niche parties rarely adjust their policy positions and are punished electorally when they do. Williams and Spoon (2015) revealed that larger parties tend to react when public opinion becomes more Euroskeptic and also that governing parties are less responsive to such changes in public opinion on this core issue of populists.

We can take away that party size, party ideology and whether a party forms part of the government or the opposition should determine its strategy towards populist challenges. Yet, the special character of German refugee policies makes it hard to formulate concrete expectations for each party regarding the adoption of populist topics. According to the literature, the CDU is predestined to incorporate populist topics and positions, since it is a large moderate right party. However, its participation in the federal government and implementation of liberal refugee policies, which is idiosyncratic in terms of party ideology, severely restrict its room to maneuver. Similar cross-pressures apply to other parties as well, since all parties except the AfD and the CSU shared a principal consensus on the most salient topic during our research period, the refugee crisis.

Research approach

Our paper thus primarily takes an exploratory approach that nonetheless adds to previous research in several regards: first, in addition to programmatic adjustments within the party system, we also analyze the populist movement Pegida. Second, prior research mostly concentrated on party manifestos which should be regarded as artifacts of strategic considerations tailored primarily towards media audiences. Party programs therefore do not necessarily reveal populist shifts in everyday political communication. Our approach using data from social media is able to cover a non-party actor in Pegida and a party that had not produced a coherent party program until 2016, the AfD. Third, the categorization of contents is unsupervised, i.e. our approach covers the universe of empirically relevant topics and is therefore more flexible than the fixed topic categories found in manifesto datasets such as the CMP. This, for instance, allows us to assess various important facets of the refugee crisis that would have been concealed within one or two higher level CMP categories.

The limitations of such a design are that we rely on topic saliences that are better suited for the analysis at a larger scale than issue positions which are more complex to operationalize.[1] Moreover, relevant co-correlates in the context of political and media systems influence the strategic considerations of political actors. Especially in light of the concurrence of AfD's and Pegida's rise with the refugee crisis, we cannot clearly attribute shifts in topic saliences by parties to one of those three factors. Nevertheless, our research design holds exogenous influences constant since actors from all analyzed political groups are similarly exposed to ongoing events and news cycles. Given their extended presence on the social network, Facebook is an ideal data source to compare political communication by Pegida, AfD and established political parties.

Methods

Data collection. For our empirical analysis, we retrieved all posts from the public Facebook pages of Pegida and German political parties. The selection of Pegida accounts relied on a list of affiliated local branches presented on the main Pegida Facebook page.[2] In addition to the main party Facebook accounts, we collected the posts from politicians affiliated with the political parties AfD, CDU, CSU, FDP, Grüne (Bündnis 90/Die Grünen), Linke (Linkspartei) and SPD at the federal level.[3] Moreover, we mined the respective public user comments and likes on the posts of the main Facebook accounts of each group. Our research period starts on the day of the opening of the Pegida Dresden account on 29 December 2014 and lasts until 17 August 2016. For the data mining, we connected to the Facebook Graph API using the R package *Rfacebook* (Barberá, 2016). The final dataset is described in Table 1.

It is noteworthy that our data do not contain posts and comments that had either been deleted by users, moderators of the political Facebook pages or by Facebook, since our data mining was conducted ex post on 17 August 2016. Therefore, the dataset depicts the curated self-presentation and as such the strategic considerations of political actors our research aims to reveal. Furthermore, we only chose the parties with a realistic chance of passing the electoral threshold of 5% required for representation in the Bundestag. Fringe parties like the Pirates or the NPD predominantly discuss niche topics and do not necessarily tailor their messages strategically in order to appeal to large shares of voters at the federal level.

Table 1. Description of the Facebook dataset.

Party	Accounts	Posts	Likes	Comments
AfD	128	68,875	14,363,982	1,865,905
CDU	180	63,057	3,772,036	1,011,581
CSU	34	13,227	4,571,095	648,375
FDP	103	38,275	7,296,749	800,281
Grüne	53	23,065	2,432,002	426,665
Linke	55	27,406	7,387,873	588,105
Pegida	25	34,282	5,318,992	850,672
SPD	172	87,115	4,739,772	613,699
Total	750	355,302	49,882,501	6,805,283

User behavior analysis. In the first step of our analysis, we concentrate on behavioral patterns of users engaging with posts created by political actors on Facebook.[4] More specifically, we are interested in the exposure of the seven main party accounts to Pegida supporters. To measure this exposure, we extracted all users that liked or commented on posts at least once.[5] We then calculated the overlaps between unique users of each party (or Pegida) in the likes and comments layers (see Equation (1)).

$$overlap_{group\ x,y} = \frac{|group\ x \cap group\ y|}{|group\ x|}. \tag{1}$$

Text analysis. In the second step of our study, we concentrate on the contents of political communication. Using the posts by Pegida, parties and politicians in the dataset described above, we analyze the extent to which the political groups discuss different topics and how these topic mixtures change over time. In order to identify the topics contained in our dataset, we employ *Latent Dirichlet Allocation* (LDA). LDA is an unsupervised Bayesian form of latent semantic analysis introduced by Blei, Ng, and Jordan (2003). Our decision to use LDA, a mixed membership model, is based on the assumption that Facebook posts can contain more than one topic per post. For the analysis of different groups' topic mixtures, we average the topic distributions of each post by each group. For analyzing how a group's topic mixture changes over time, we average the topic distributions of the group's posts for each day.

To reduce the linguistic complexity of the posts, we applied the following preprocessing steps: First, we removed German stopwords, links, words shorter than three characters, as well as very frequent words and words occurring in less than 10 posts. Next, we removed the names of sitting members of the German parliament. Finally, we removed all posts that had less than five words remaining. All of these steps serve the goal of obtaining interpretable topics depicting the political issues discussed in the data (Grimmer & Stewart, 2013). After preprocessing, the dataset consists of 244,237 posts, with a vocabulary of 50,166 unique terms.

We then trained LDA with different levels of granularity (50 and 100 topics) on the preprocessed corpus. For training, we used the Collapsed Variational Bayes inference schema (Teh, Newman, & Welling, 2006), as implemented by Ramage and Rosen (2010) with 200 iterations. The model priors α and β were set to 0.1 and 0.01, respectively. These low prior values reflect our beliefs that Facebook posts tend to cover few topics in one post (as opposed to many different topics) and that the covered topics contain relatively few, specific words. The two separate model runs resulted in similar topic groupings, yet with different levels of granularity. The most important criterion when evaluating topic model outputs should be their substantive fit in the context of a specific research question (Grimmer & Stewart, 2013, p. 286). In that regard, the configuration with 100 topics produced the most appropriate topics which we will evaluate and use in the empirical analysis.

Results and discussion

User behavior analysis

To answer our first research question on overlaps between audiences of Pegida and political parties, we calculated the intersections of unique users in the likes and comments layers (see Equation (1)). The underlying assumption is that a high exposure to Pegida supporters creates incentives to address topics popular with these users. Figure 1 displays the fraction of the unique user base of party *x* (left column) also having liked or commented at least one post by party *y* (bottom row). Therefore, the values for party pairs (and Pegida-party pairs) differ in the boxes above and beneath the grey diagonal.

Several findings in Figure 1, Panel A stand out. First, 33% of Pegida likers, 79,333 users, liked contents on the AfD site at least once. This is the highest overlap of all party pairs. In return, of AfD unique likers, 21% liked Pegida contents, making this party pair the one with the highest reciprocal affinity. Second, of CSU unique users, considerably shares also liked AfD contents (21%) and Pegida contents (11%). In contrast, only 6% of CSU sympathizers also liked a post by the sister party CDU, whereas it is 22% the other way round. Furthermore, the CSU is the only established party with a certain appeal to Pegida likers (13%) and AfD likers (15%). Third, the overlaps between

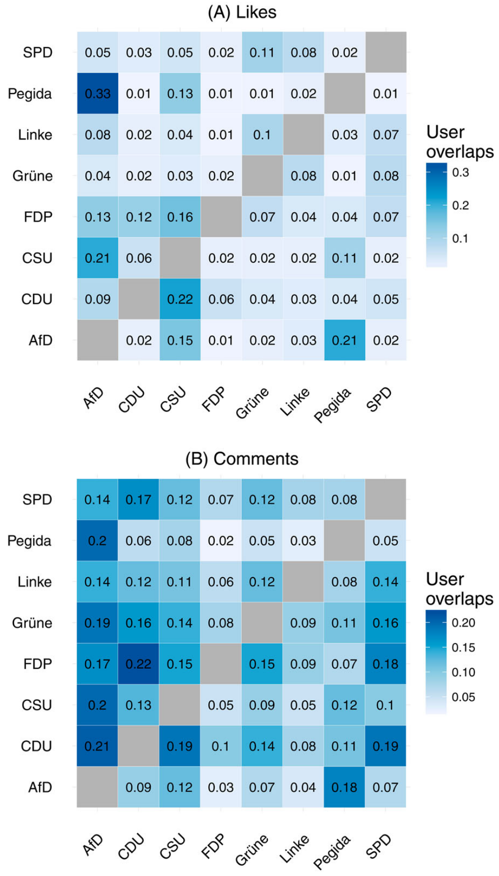


Figure 1. Overlaps in unique users per group.

other party pairs are mostly negligible. Only FDP users allocate likes widely to other parties. Besides that, we observe an ideological sorting along party lines in the likes of mainstream political parties. The clearest conglomeration of user bases takes place between Pegida, AfD and CSU. The common denominator between these three groups is their position furthest right on the political spectrum and their opposition to liberal refugee policies.

In Panel B of Figure 1, we see a much more heterogeneous picture. Pegida and AfD still have the highest reciprocity in terms of users commenting on both sites. Yet, their users and those of all other parties distribute comments much more evenly than likes.

The differences between the likes and comments layers indicate that these conventions are used for different purposes and therefore have diverging social meanings when regarded as aggregate counts. Likes can be regarded as a sign of political support which Pegida users mostly attribute to posts of AfD, but also to the CSU which criticized the governments' refugee policy (in which the CSU is participating, ironically). Comments, in contrast, are oftentimes also used to voice criticism and therefore allocated rather evenly across the political spectrum. The diverging use of likes and comments is in line with findings from Twitter, where retweets are a stronger predictor of partisan homophily than @-mentions (Lietz, Wagner, Bleier, & Strohmaier, 2014).

The results show that the supporter bases of Pegida and AfD are the most similar ones across the German political spectrum. This could influence political communication by the AfD and, reciprocally, populist communication by the AfD should attract Pegida supporters. Other German parties besides the CSU are exposed much less frequently to users also active on the Pegida site.

Topics addressed by Pegida and political parties

By implication, the previous analysis suggests that political communication by Pegida and AfD should be rather similar as well, since it resonates with an identical group of politically interested people of considerable size. We take to the content level of political actors' posts to investigate this systematically. For this, we train the LDA topic model on the full corpus of posts by Pegida, parties and individual politicians. As described in the section 'Methods', we use the model with 100 topics.

To narrow down the scale to politically interpretable and thus substantively relevant topics, three of the authors independently coded the model outputs as relevant topics on policies or contemporary events, or of no substantive interest (with an inter-rater reliability of Fleiss' Kappa $= 0.706$, $p < .000$). This means that 'stopword topics' containing Facebook-specific language such as 'like, follow, share' or parliamentary procedural topics containing 'vote, debate, speaker' were dropped, but also topics on constituency service which, while being interpretable, cover procedural instead of substantive issues and are therefore of no relevance here. We also excluded party-specific topics with predominantly organizational information and unique language only used by a particular group. This procedure left us with 46 topics of substantive interest out of the original 100. The authors independently assigned titles based on the top scoring words for each topic in the appendix and decided on the few ambiguous cases consensually. The model identifies a mix of generally relevant policy fields but also topics more specific to our research period like the Euro or refugee crises.

For our subsequent content analysis, we calculated the average topic probabilities for each group in the 46 topics (Figure 2). The higher its share in a topic, the more heavily a group referred to a topic in their Facebook posts. It is important to note that the *y*-axes are flexible which means that each plot has an individual value range. From the perspective of the literature on party competition, the results are ambiguous. On the one hand, there is a skew towards one or few parties in many of the topics which is exactly what issue saliency theory predicts (Budge & Farlie, 1983). A lot of the variation can be explained by the diverging core competencies parties have, e.g. Grüne overemphasize *Energy/climate policy* while SPD and Linke frequently talk about *Social policy – Unions*. On the other hand, various topics like *Terror attacks in Europe* are well balanced with similar topic shares by several groups across the political spectrum. This reveals the limitations of an approach relying exclusively on topic salience. In some topics, the positional competition model which distinguishes between different political preferences on identical topics (Dolezal et al., 2014) is more applicable.

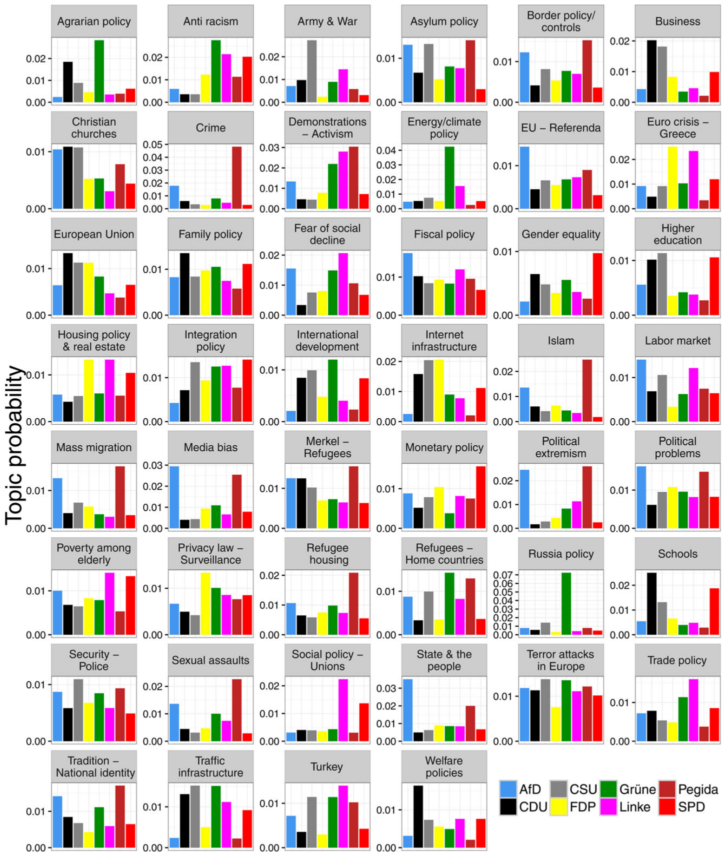
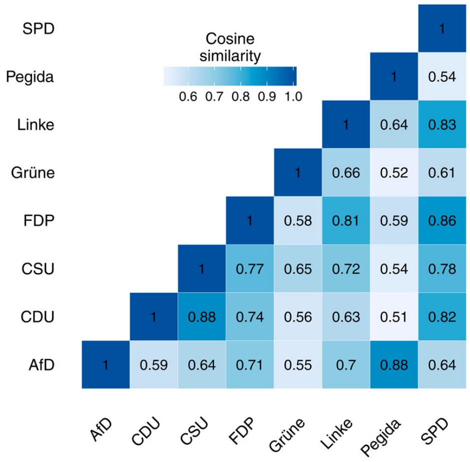


Figure 2. Topics addressed by Pegida and political parties.

To shed light on similarities between groups, we statistically compare the vectors of party-specific topic probabilities across the 46 topics. For this, we calculate the cosine similarities of these distributions for all party pairs.[6] Figure 3 shows that with a cosine similarity of 0.88, Pegida and AfD have the most similar distributions (together with the sister parties CDU and CSU).[7] As it is to be expected, the coalition partners at the federal level CDU and SPD also discuss similar issues. Counterintuitively, the FDP vector has a high congruence with the topic vectors of Linke and SPD, which are both parties from the left. It seems that on Facebook, the multifaceted FDP emphasizes its progressive side, e.g. in the topic *Privacy law – Surveillance*, but also has high probabilities jointly with leftist parties in topics like *Euro crisis – Greece* or *Housing policy and real estate* on which the FDP certainly proposes diverging positions.

Topics emphasized by populists

The high cosine similarity between Pegida and AfD shows that they not only attract similar users but also discuss similar political issues. To assess the extent to which the identified topics should be considered as typical items on a populist agenda, external judgments are needed. For this, we set up a survey on the crowdsourcing platform CrowdFlower.[8] While it is not feasible to train non-experts sufficiently to code a concept like populism that is even disputed in the academic literature (Reinemann et al., 2016), crowdworkers provide a more diverse set of opinions than a small group of authors and research assistants. Moreover, crowd tasks scale up well, i.e. many respondents can be recruited to judge 46 topics in a swift and affordable way. This survey is probably not representative of the German population. However, German crowdworkers are exposed to media coverage on and political communication by populist actors. In essence, their ratings represent the extent to which they perceive different topics as salient on a populist agenda.



	AfD	CDU	CSU	FDP	Grüne	Linke	Pegida	SPD
SPD								1
Pegida							1	0.54
Linke						1	0.64	0.83
Grüne					1	0.66	0.52	0.61
FDP				1	0.58	0.81	0.59	0.86
CSU			1	0.77	0.65	0.72	0.54	0.78
CDU		1	0.88	0.74	0.56	0.63	0.51	0.82
AfD	1	0.59	0.64	0.71	0.55	0.7	0.88	0.64

Figure 3. Cosine similarities between topic distributions of Pegida and political parties.

German crowdworkers were first provided the three criteria of populist communication defined by Reinemann et al. (2016) (see section ‘Related research and expectations’). Then they were asked about their opinion on the extent to which the 10 most typical words per topic (appendix) represent populist communication. For this, they had to rate the keywords in each topic on a scale from 1 (not at all populist) to 7 (clearly populist). Three attention checks were included in the survey to detect and filter out spammers. Of the 150 crowdworkers we surveyed, 107 passed all the attention checks. Only the ratings by the latter respondents were kept in our analysis.

The average *populism rating* per topic is listed in Table 2. The aggregated allocation of topics to ratings is almost normally distributed ($N = 46$, mean = 3.86, median = 3.78).

Table 2. Crowdworker ratings per topic.

Topic	Populism rating	Standard deviation
Media bias	5.53	1.58
Border policy/controls	5.29	1.49
Political extremism	5.27	1.51
Islam	5.19	1.55
Turkey	5.00	1.69
Asylum policy	4.93	1.80
Sexual assaults	4.84	1.78
Mass migration	4.77	1.61
Refugee housing	4.70	1.77
Merkel – Refugees	4.70	1.84
EU – Referenda	4.64	1.72
Anti racism	4.59	1.95
Fear of social decline	4.56	1.63
Poverty among elderly	4.53	1.76
Trade policy	4.46	1.70
Russia policy	4.42	1.57
Demonstrations – Activism	4.28	1.83
Integration policy	4.23	1.94
Terror attacks in Europe	4.12	2.01
Army & War	3.99	1.86
Euro crisis – Greece	3.93	1.70
Privacy law – Surveillance	3.92	1.65
Refugees – Home countries	3.80	1.86
Fiscal policy	3.78	1.74
Crime	3.77	1.75
Security – Police	3.70	1.62
State and the people	3.66	1.64
Housing policy and real estate	3.63	1.74
Monetary policy	3.61	1.68
Labor market	3.57	1.87
International development	3.54	1.68
Social policy – Unions	3.26	1.74
Christian churches	3.20	1.81
Tradition – National identity	3.16	1.79
Traffic infrastructure	3.04	1.76
Agrarian policy	3.01	1.79
Welfare policies	2.97	1.82
Political problems	2.93	1.71
European Union	2.92	1.76
Energy/climate policy	2.79	1.64
Internet infrastructure	2.79	1.68
Family policy	2.65	1.73
Gender equality	2.64	1.61
Business	2.55	1.59
Higher education	2.47	1.62
Schools	2.30	1.67

While the standard deviations indicate considerable disagreement between respondents, the topic means were similar in a repetition of the survey (30 respondents; Spearman correlations between both surveys: $\rho = 0.92$). In order to evaluate these ratings, we qualitatively assess the topics in the top ranks according to the three pillars of populist communication defined by Reinemann et al. (2016). To give the reader insights how the topics manifest themselves in messages, we listed the five Facebook posts with the highest probability in the top 10 topics in the online appendix.

In terms of criticism of elites, the topic *Media bias* stands out, as demonstrated by negative attributions like 'truth', 'propaganda' or 'lügenpresse' (Media liars), one of the most salient slogans of Pegida (see appendix). A substantial amount of criticism towards elites is also contained in the topics *EU – Referenda* and *Merkel – Refugees*. However, Figure 2 shows that even though Pegida and AfD share the highest probabilities, other parties also contribute to the latter topics, which implies that the interpretation of these topics by populists is contested by differing partisan positions.

Multiple topics focus on issues related to perceived out-groups. Most evidently, aspects of refugee policy such as *Asylum policy*, *Border policy/controls*, *Islam*, *Mass migration* and *Refugee housing* are central aspects in populist communication. The vocabulary on *Political extremism* indicates that it serves the purpose to define 'out groups' by circumscribing an own core of supporters from groups at the other end of the political spectrum perceived as being violent (e.g. 'Antifa', a left wing anti-racist group). Moreover, crowdworkers attached a high populist rating to the wordlists in the topic *Sexual assaults*, which clearly refers to the incidents on New Years' Eve 2015/2016 ('Cologne') when women were perpetrated predominantly by men from the Middle East and Northern Africa. In all of these topics, Pegida and AfD again share the highest probabilities (Figure 2).

Three topics are particularly emphasized by leftist parties. *Fear of social decline* and *Poverty among elderly* address latent fears of a social decline and are not only emphasized by Grüne, Linke and SPD but also the right-wing AfD. Meanwhile, *Trade policy* concerns issues typically made salient by the political left such as international treaties like TTIP or CETA or the drilling technique fracking. One defining feature of our minimalist conceptualization of populism as a 'thin ideology' (Mudde, 2004) is its openness to diverse ideologies. Accordingly, leftist actors like Bernie Sanders, Syriza in Greece, the Five Star movement in Italy and Podemos in Spain have all been labeled as populists by media observers and academics (Mudde, 2015).

Two of the ratings in the top 15 are rather ambiguous. The deal concerning refugees with *Turkey* is frequently discussed by populists, but at the same time other parties emphasize different aspects of Turkey policies. The topic *Anti racism* is primarily devoted to mobilize against right-wing tendencies and was perceived by crowdworkers as a form of populism at the other end of the political spectrum. Both results point towards the limitations of the necessarily parsimonious information provided to crowdworkers and the need to further distinguish different aspects and positions parties emphasize on identical topics.

Within the top ranks, there are also no clear references to the arguably most important pillar of populism, the promise to advocate for the pure interests of the people. But as noted by Reinemann et al. (2016), the communicative construction of the people can also be made implicitly by contrasting this idealized homogeneous body to the problems attached to elites and out-groups. German populists predominantly prefer frames attacking perceived outsiders rather than making references to the *Volk*, a restraint which might be related to the extensive use of this notion in the propaganda of the Third Reich. Yet, recent public statements of AfD party leaders like the initiative of Frauke Petry to exculpate the term *völkisch* of its troublesome past indicates that these communicative taboos are not off-limits to populists anymore.

Temporal patterns in topic salience

In our final analysis, we analyze which topics parties emphasize over time, in particular topics with a populist appeal. For this, we use the crowdsourced populism ratings to calculate a *topic salience weighted by populism rating* (TSPR) for each political group x per day t via the equation

$$TSPR_{xt} = \sum_{k=1}^{K} salience_{kxt} * rating_k, \tag{2}$$

where $K = 46$ is the number of topics. In order to identify shifts over time, we use these daily values as input to fit LOESS regressions per group. The regressions predict the daily values by taking into account the neighboring data points in the time series. This data fitting technique smooths the considerable daily volatility in the raw time series data, removes seasonality effects and facilitates the identification of trends.

Several patterns can be observed in Figure 4. First, Pegida and AfD have the highest topic salience weighted by populism rating. Their time series reached peaks during the height of the refugee crisis in the second half of 2015. Afterwards, they seem to have shifted their focus to other topics, however, with an upward trend again since April 2016.

Second, the two biggest parties CDU and SPD which form a coalition in the federal government have the lowest values on this scale. As the refugee crisis unfolded, they addressed related issues, yet deemphasized them again in 2016.

Third, the time series for the smaller parties CSU, FDP, Linke and Grüne are quite volatile. When inspecting the individual time series of parties in each topic, it becomes clear that FDP and Linke emphasized topics with a higher populism weight mostly as a reaction to external events like the increasing influx of refugees in 2015 or the events from New Years' Eve 2015/2016 in Cologne. Topic salience does not reveal their positions on these issues, yet several of their leading politicians, for instance, Sarah Wagenknecht (Linke) and Christian Lindner (FDP), publicly criticized the refugee policies by Angela Merkel's government and both have been accused of flirting with populist stances.

Furthermore, the aggregated time series mask heterogeneous topical foci of parties. The Grüne and Linke are also prominent here because they emphasize their core issues like *Trade policy* or *Fear of social decline*. Yet, the Grüne also put a special focus on *Asylum policy*, which implies that the party actively contests the interpretation of the topic by populists, e.g. in the debate on deportations of asylum seekers. The CSU, on the other hand, increasingly talked about refugee policies by putting a particular focus on *Asylum policy*, *Border policy/controls*, and *Mass migration*.

The two bigger governing parties clearly deemphasized topics that are typically stressed by populists, in particular on refugee and migration policies. The marginalization of these issues by CDU and SPD, on which their performance was rated very critically by the public (infratest dimap, 2016), support core assumptions of the original saliency theory in party

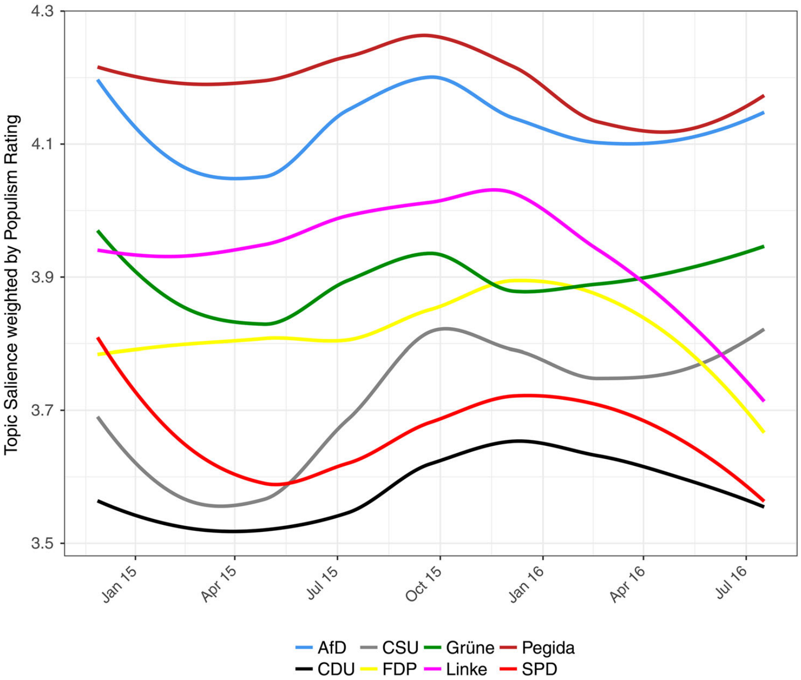


Figure 4. Topic salience weighted by populism rating over time.

competition (Budge & Farlie, 1983). However, the findings need to be complemented by more complex analyses of party positions. Especially, the results for smaller parties are less consistent but rather point to positional differences (Meguid, 2005), for instance, 'accommodative strategies' by the CSU and 'adversarial strategies' by the Grüne. The strategies how to address populist challenges can principally be located between these two poles, but certainly vary between parties as well as over time within parties.

Although generally at a high level, there is also significant temporal volatility in populism weighted topic salience by Pegida and AfD. Especially the drop off in the beginning of 2016 is of interest. Did Pegida and AfD concentrate on more moderate issues after the influx of refugees narrowed down? In order to answer this we will look at several of the topics in which Pegida and AfD share high probabilities (Figure 2).

Figure 5 shows that AfD as well as Pegida have adjusted to shifts in public attention and identified new salient topics to which they can attach their populist message. The AfD increasingly discusses *EU – Referenda* focusing on the Brexit and referenda in other European countries. Pegida has clearly stressed the topics *Crime* and *Sexual assaults* since the New Years' Eve 2015/2016.

It is also worth to look at the two topics in which Pegida and AfD share high probabilities not rated as particularly populist by crowdworkers because their use by populists only becomes apparent through their context in messages (see online appendix). First, when

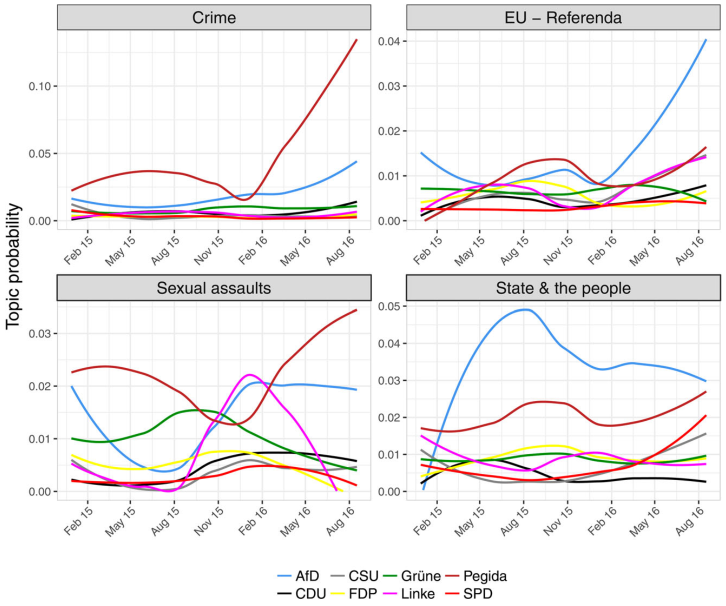


Figure 5. Topics characteristic for Pegida and AfD.

looking at posts regarding *Crime*, it becomes clear that AfD and Pegida relate criminal activity to the presence of refugees in the country. The use of this frame targeting the sense of security of audiences can be compared to 'personalized frames' of leftist movements (Bennett et al., 2014) and fits well with the populist aspiration to protect the people from dangers instigated by out-groups. Second, Pegida and AfD attach particular relevance to the *State and the people*. This topic is the clearest manifestation that both also discuss general questions regarding the role of the *Volk* within the polity.

Conclusion

This paper set out to analyze online communication by populist actors and established political parties in Germany. The results show that Pegida and AfD appeal to similar target groups by emphasizing rather unique topics in their communication on social media. While party leaders repeatedly distanced the AfD from more radical right-wing groups and Pegida in particular, our findings challenge the self-presentation of the AfD as a party of the political center. These results add to the literature on populism in Western democracies that has so far exclusively focused on populist challengers in the form of political parties. The analysis of Pegida's Facebook activity also demonstrated that not only leftist social movements are adept in using social media and personalized frames (Bennett, 2012), but also counterparts at the opposite end of the ideological spectrum. In a counterfactual situation without the emergence of Pegida, which revealed the substantial resonance of right-wing positions in German society, the more nationalist forces in the AfD might not have prevailed in the internal party struggle during 2015. Since then, the AfD has firmly established itself as the melting pot of populism in the German party system.

Coupled together, populist actors in Germany have a considerable audience that is constantly mobilized against the *Altparteien* and perceived out-groups. We analyzed whether established parties have adapted to these pressures by addressing similar topics in their political communication. Our analysis over time revealed limited evidence for an increasing emphasis of core populist topics in messages by established parties. The parties on the political left, Grüne and Linke emphasized some of their core issues like trade policy which can be subsumed under a minimalist definition of populism. The CSU increasingly addressed topics emphasized by populists but still at a significantly lower level than Pegida and AfD. However, whether these fluctuations are causally related to right-wing populism is beyond the scope of this study. These shifts, which we interpret tentatively, are in line with previous inconclusive results regarding a programmatic contagion in the mainstream party system induced by right-wing challengers (Bale et al., 2010; Rooduijn et al., 2014).

From the perspective of the literature on party competition, the most interesting finding is that the conservative party CDU deemphasized populist topics which runs contrary to previous studies (Abou-Chadi, 2016; Bale, 2003). German politics during our research period should be regarded as a special case since a 'grand coalition' consensually implemented liberal refugee policies. CDU, SPD and all established parties – with the exception being the CSU – deemphasized related issues, which nonetheless remained salient due to external events. This created a vacuum that populists exploited. In other countries, such a consensus across parties is not to be expected and especially conservative parties position themselves more to the right than Angela Merkel's CDU. In order to test to which extent these results hold in other contexts, our methodology could be applied to further cases, since data can be gathered ex post via the Facebook Graph API. In general, the approach can be used for the analysis of substantive issues other than populism and in various disciplines, whenever a constellation is present in which new groups or ideas enter an established social system.

We also want to address the limitations of this study. Several uncertainties accompanied the data collection process, since we could not retrieve posts that had been deleted by holders of political accounts, users or Facebook moderators. Furthermore, a systematic coding at the level of posts may be more accurate than at the topic level, but is on the other hand more complex to implement in terms of the anonymization required to conceal the sender, costs, time and personnel, even when deploying crowdworkers. Moreover, the topics identified by the unsupervised LDA model are naturally dependent on the underlying data from a specific research period. A promising direction for future work could be to apply a polylingual labeled topic model (Posch, Bleier, Schaer, & Strohmaier, 2015) which is able to incorporate both predefined labels related to populist rhetoric and additional information about the posts such as the language characteristics of user comments. Considering the variety of methodological opportunities, quantitative text analysis holds great promise to improve the analysis of populist communication going beyond the infrequently published party manifestos.

Notes

1. In further research, we will extend our approach to positional competition.

2. There are indications that the newly created Pegida branches are more radical than the Dresden chapter (Vorländer et al., 2016, p. 69). However, since the accounts are listed and liked by the main Pegida site, they are clearly regarded as part of the movement by its leaders.
3. We relied on several data sources. We thank Martin Fuchs and his website Pluragraph for providing us with lists of the social media accounts of sitting parliamentarians in the federal parliament (*Bundestag*) and leading politicians of the non parliamentarian parties AfD and FDP. The list of AfD politicians also contains the candidates for the German federal election 2013 (Kaczmirek & Mayr, 2015) except the ones who have left the AfD and joined ALFA, the new party of AfD founder Lucke. To increase the share of messages coming from official party accounts, which we assume communicate more strategically than individual politicians, we included the accounts of the parties in the federal states.
4. The Facebook Graph API does not provide information on who likes Facebook pages themselves.
5. Pegida's account was deleted by Facebook on 22 July because of 'instances of hate speech'. Therefore we could only conduct the behavioral analysis based on data retrieved in a previous data crawl. Figure 1 is consequently based on all unique users engaging with a post by one of the eight main accounts before February 20, 2016.
6. The results are robust when using the Jensen-Shannon divergence as a distance metric.
7. As a robustness test, we compared the cosine similarities between all groups in the two models with 50 and 100 topics resulting in a Spearman rank correlation of $\rho = 0.84$. The party specific topic distributions are therefore very similar independent of the number of topics.
8. https://www.crowdflower.com

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Sebastian Stier is a postdoctoral researcher in the Department Computational Social Science at GESIS – Leibniz Institute for the Social Sciences in Cologne. He studies the role of the Internet in political communication relying on theoretical approaches and methods from Comparative Politics, Communication Sciences and Computational Social Science [email: sebastian.stier@gesis.org].

Lisa Posch is a doctoral student in the Department Computational Social Science at GESIS. Her research is focused on human intelligence tasks and topic models, and their application within the Computational Social Science domain [email: lisa.posch@gesis.org].

Arnim Bleier is a postdoctoral researcher in the Department Computational Social Science at GESIS. His research interests are in the field of Computational Social Science. In collaboration with social scientists, he develops Bayesian models for the content, structure and dynamics of social phenomena [email: arnim.bleier@gesis.org].

Markus Strohmaier is a Full Professor of Web-Science at the Faculty of Computer Science at the University of Koblenz-Landau and scientific director of the Department Computational Social Science at GESIS. His main research interests include Web-Science, Computational Social Science and Data Science [email: markus.strohmaier@gesis.org].

ORCID

Sebastian Stier http://orcid.org/0000-0002-1217-5778

References

Abou-Chadi, T. (2016). Niche party success and mainstream party policy shifts – how green and radical right parties differ in their impact. *British Journal of Political Science*, *46*(02), 417–436.

Adams, J., Clark, M., Ezrow, L., & Glasgow, G. (2006). Are niche parties fundamentally different from mainstream parties? The causes and the electoral consequences of Western European parties' policy shifts, 1976–1998. *American Journal of Political Science*, *50*(3), 513–529.

Arzheimer, K. (2015). The AfD: Finally a successful right-wing populist eurosceptic party for Germany? *West European Politics*, *38*(3), 535–556.

Bale, T., Green-Pedersen, C., Krouwel, A., Luther, K. R., & Sitter, N. (2010). If you can't beat them, join them? Explaining social democratic responses to the challenge from the populist radical right in Western Europe. *Political Studies*, *58*(3), 410–426.

Bale, T. (2003). Cinderella and her ugly sisters: The mainstream and extreme right in Europe's bipolarising party systems. *West European Politics*, *26*(3), 67–90.

Barberá, P. (2016). Package 'Rfacebook'. CRAN.

Bennett, W. L., Segerberg, A., & Walker, S. (2014). Organization in the crowd: Peer production in large-scale networked protests. *Information, Communication & Society*, *17*(2), 232–260.

Bennett, W. L. (2012). The personalization of politics: Political identity, social media, and changing patterns of participation. *The ANNALS of the American Academy of Political and Social Science*, *644*(1), 20–39.

Berbuir, N., Lewandowsky, M., & Siri, J. (2015). The AfD and its sympathisers: Finally a right-wing populist movement in Germany?. *German Politics*, *24*(2), 154–178.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Budge, I., & Farlie, D. (1983). *Explaining and predicting elections: Issue effects and party strategies in twenty-three democracies*. London: George Allen & Unwin.

CNN (2016). *Occupy Wall Street rises up for Sanders*. Retrieved from http://edition.cnn.com/2016/04/13/politics/occupy-wall-street-bernie-sanders-new-york-primary.

Chadwick, A. (2013). *The hybrid media system: Politics and power*. Oxford: Oxford University Press.

Die Welt (2015). Vorauseilende Pegida-Schelte hilft nur der AfD. Retrieved from http://www.welt.de/debatte/kommentare/article136261979/Vorauseilende-Pegida-Schelte-hilft-nur-der-AfD.html.

Dolezal, M., Ennser-Jedenastik, L., Müller, W. C., & Winkler, A. K. (2014). How parties compete for votes: A test of saliency theory. *European Journal of Political Research*, *53*(1), 57–76.

Downs, A. (1957). *An economic theory of democracy*. New York, NY: Harper & Row.

Engesser, S., Ernst, N., Esser, F., & Büchel, F. (2016). Populism and social media: How politicians spread a fragmented ideology. *Information, Communication & Society*, *20*(8), 1109–1126.

Frees, B., & Koch, W. (2015). Internetnutzung: Frequenz und Vielfalt nehmen in allen Altersgruppen zu. *Media Perspektiven*, *9*, 366–377.

González-Bailón, S., & Wang, N. (2016). Networked discontent: The anatomy of protest campaigns in social media. *Social Networks*, *44*(1), 95–104.

Green-Pedersen, C. (2007). The growing importance of issue competition: The changing nature of party competition in Western Europe. *Political Studies*, *55*(3), 607–628.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.

Hein, T. (2017). Pegida als leerer Signifikant, Spiegel und Projektionsfläche – eine Einleitung. In T. Hein (Ed.), *Pegida als Spiegel und Projektionsfläche* (pp. 1–31). Wiesbaden: Springer VS.

infratest dimap (2016). ARD-DeutschlandTREND Februar 2016. Retrieved from http://www.infratest-dimap.de/fileadmin/user_upload/dt1602_bericht.pdf.

Kaczmirek, L., & Mayr, P. (2015). German Bundestag Elections 2013: Twitter usage by electoral candidates.

Korsch, F. (2016). Natürliche Verbündete?. *Die Pegida-Debatte in der AfD zwischen Anziehung und Ablehnung* (pp. 111–134). Wiesbaden: Springer Fachmedien Wiesbaden.

Lietz, H., Wagner, C., Bleier, A., & Strohmaier, M. (2014). When politicians talk: Assessing online conversational practices of political parties on Twitter. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media* (pp. 285–294). Palo Alto, CA: AAAI Press.

McAdam, D., & Tarrow, S. (2010). Ballots and barricades: On the reciprocal relationship between elections and social movements. *Perspectives on Politics*, *8*(2), 529–542.

Meguid, B. M. (2005). Competition between unequals: The role of mainstream party strategy in niche party success. *American Political Science Review*, *99*(03), 347–359.

Mudde, C. (2004). The populist Zeitgeist. *Government and Opposition*, *39*(4), 542–563.

Mudde, C. (2015). The problem with populism. Retrieved from https://www.theguardian.com/commentisfree/2015/feb/17/problem-populism-syriza-podemos-dark-side-europe.

Pegida (2015). Zehn Thesen in Dresden angeschlagen – Dafür geht PEGIDA auf die Straße. Retrieved from https://pegidaoffiziell.wordpress.com/2015/02/16/zehn-thesen-in-dresden-angeschlagen-dafur-geht-pegida-auf-die-strase.

Posch, L., Bleier, A., Schaer, P., & Strohmaier, M. (2015). The polylingual labeled topic model. In *Joint German/Austrian conference on artificial intelligence* (pp. 295–301). Cham: Springer.

Ramage, D., & Rosen, E. (2010). *Stanford topic modeling toolbox*. Retrieved from http://nlp.stanford.edu/software/tmt/tmt-0.4/.

Reinemann, C., Aalberg, T., Esser, F., Strömbäck, J., & de Vreese, C. H. (2016). Populist political communication. Toward a model of its causes, forms, and effects. In T. Aalberg, F. Esser, C. Reinemann, J. Stromback, & C. De Vreese (Eds.), *Populist political communication in Europe* (pp. 12–25). New York, NY: Routledge.

Rooduijn, M., de Lange, S. L., & van der Brug, W. (2014). A populist Zeitgeist? Programmatic contagion by populist parties in Western Europe. *Party Politics*, *20*(4), 563–575.

Teh, Y. W., Newman, D., & Welling, M. (2006). A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation. In *Proceedings of the 20th Annual Conference on Neural Information Processing* (pp. 1353–1360). Cambridge, MA: MIT Press.

Vorländer, H., Herold, M., & Schäller, S. (2016). *PEGIDA: Entwicklung, Zusammensetzung und Deutung einer Empörungsbewegung*. Wiesbaden: Springer VS.

Williams, C., & Spoon, J.-J. (2015). Differentiated party response: The effect of Euroskeptic public opinion on party positions. *European Union Politics*, *16*(2), 176–193.

Appendix. Top words in each topic

Agrarian policy:
landwirtschaft, ernährung, verbraucher, lebensmittel, tiere, glyphosat, landwirte
(agriculture, food, consumer, groceries, animals, glyphosate, farmers)
Anti racism:
zeichen, rassismus, setzen, gewalt, hass, toleranz, zeigen
(sign, racism, put, violence, hate, tolerance, show)
Army & War:
bundeswehr, einsatz, soldaten, krieg, hand, syrien, waffen
(bundeswehr, mission, soldiers, war, hand, syria, arms)
Asylum policy:
asylbewerber, asyl, sicheren, asylverfahren, asylbewerbern, asylrecht, abschiebung
(asylum seeker, asylum, secure, asylum procedure, asylum seekers, asylum law, deportation)
Border policy/controls:
grenzen, grenze, österreich, grenzkontrollen, schützen, regierung, kontrolle
(borders, border, austria, border controls, protect, government, control)
Business:
unternehmen, region, wirtschaft, mitarbeiter, firma, geschäftsführer, gmbh
(companies, region, economy, employees, company, manager, gmbh)
Christian churches:
kirche, schaut, christen, roten, evangelischen, kirchen, katholischen
(church, look, christians, red, protestant, churches, catholic)
Crime:
täter, mehrere, verletzt, männer, polizisten, laut, nacht
(offender, multiple, injured, men, policemen, loud, night)
Demonstrations - Activism:
straße, demo, demonstration, samstag, platz, kundgebung, teilnehmer
(street, demo, demonstration, saturday, square, rally, participants)
Energy/climate policy:
energiewende, energie, klimaschutz, umwelt, energien, strom, erneuerbaren
(energy transition, energy, climate protection, environment, energies, electricity, renewable)
EU - Referenda:
entscheidung, frankreich, brexit, großbritannien, europas, briten, entschieden
(decision, france, brexit, great britain, europe's, brits, decided)
Euro crisis - Greece:
regierung, griechischen, griechische, griechenlands, reformen, tsipras, verhandlungen
(government, greek, greek, greece's, reforms, tsipras, negotiations)
European Union:
europäischen, union, europäische, brüssel, kommission, gemeinsame, parlament
(european, union, european, brussels, commission, common, parliament)
Family policy:
familie, eltern, kindern, familien, bildung, schulen, kind
(family, parents, children, families, education, schools, child)
Fear of social decline:
folgen, angst, bevölkerung, druck, zeigen, setzt, führt
(consequences, fear, population, pressure, show, put, lead)
Fiscal policy:
geld, milliarden, millionen, kosten, steuerzahler, zahlen, haushalt
(money, billions, millions, costs, taxpayer, numbers, budget)
Gender equality:
foto, phototheknet, männer, amt, schwesig, brandenburger, manuela
(photo, phototheknet, men, office, schwesig, brandenburg, manuela)
Higher education:
bildung, prof, forschung, ausbildung, universität, wissenschaft, oldenburg
(education, prof, research, qualification, university, science, oldenburg)
Housing policy & real estate:
fordert, wohnungen, wohnen, wohnraum, fordern, forderung, sozialen
(demand, apartments, live, housing space, demand, claim, social)
Integration policy:
integration, flüchtlingen, helfen, ort, aufnahme, schutz, flucht
(integration, refugees, help, location, accommodation, shelter, escape)
International development:
zusammenarbeit, entwicklung, menschenrechte, botschafter, nationen, internationalen
(cooperation, development, human rights, ambassador, nations, international)

Internet infrastructure:
wirtschaft, chancen, digitalisierung, raum, bildung, digitale, ländlichen
(economy, chances, digitalization, space, education, digital, rural)
Islam:
islam, muslime, religion, islamischen, muslimischen, islamisierung, staat
(islam, muslims, religion, islamic, muslim, islamization, state)
Labor market:
mindestlohn, zahl, zahlen, prozent, arbeitsmarkt, millionen, deutlich
(minimum wage, number, numbers, percentage, labor market, millions, distinct)
Mass migration:
zuwanderung, einwanderung, migranten, bevölkerung, regeln, integration, gesetze
(immigration, immigration, migrants, population, rules, integration, laws)
Media bias:
medien, politiker, presse, lügenpresse, wahrheit, berichterstattung, journalisten
(media, politicians, press, lying press, truth, coverage, journalists)
Merkel - Refugees:
kanzlerin, bundeskanzlerin, flüchtlingspolitik, spricht, flüchtlingskrise, merkels, worte
(chancellor, chancellor, refugee policy, talks, refugee crisis, merkel's, words)
Monetary policy:
ezb, banken, urteil, bargeld, bank, abschaffung, bundesverfassungsgericht
(ecb, banks, verdict, cash, bank, abolishment, constitutional court)
Political extremism:
gewalt, linken, antifa, rechts, kampf, angriffe, politisch
(violence, leftists, antifa, rightists, fight, attacks, political)
Political problems:
probleme, verantwortung, problem, lösung, situation, lage, handeln
(problems, responsibility, problem, solution, situation, condition, act)
Poverty among elderly:
rente, armut, soziale, hartz, einkommen, steuern, erhöhung
(pension, poverty, social, hartz, income, taxes, increase)
Privacy law - Surveillance:
bild, maas, vorratsdatenspeicherung, links, minister, rechts, teilt
(bild, maas, data preservation, left, minister, right, shares)
Refugee housing:
asylbewerber, derzeit, unterbringung, flüchtlingen, pro, unterkunft, untergebracht
(asylum seeker, currently, accomodation, refugees, pro, shelter, accommodated)
Refugees - Home countries:
syrien, osten, mittelmeer, millionen, nahen, afrika, irak
(syria, east, mediterranean, millions, middle, africa, iraq)
Russia policy:
russland, ukraine, usa, polen, nato, syrien, russischen
(russia, ukraine, usa, poland, nato, syria, russian)
Schools:
schüler, schülerinnen, schule, klasse, bad, schülern, schulen
(pupils, pupils, school, class, bath, pupils, schools)
Security - Police:
sicherheit, polizisten, innere, schutz, stellen, personal, öffentlichen
(security, policemen, internal, protection, positions, staff, public)
Sexual assaults:
köln, straftaten, kölner, täter, gewalt, übergriffe, sexuelle
(cologne, crime, cologne's, perpetrator, violence, assaults, sexual)
Social policy - Unions:
beschäftigten, gewerkschaften, leiharbeit, dgb, arbeitgeber, arbeitsbedingungen, sozial
(employees, unions, contract work, dgb, employers, labor conditions, social)
State & the people:
art, staat, nämlich, volk, völlig, weise, eher
(manner, state, namely, people, entirely, way, rather)
Terror attacks in Europe:
paris, terror, opfer, gedanken, angehörigen, anschlag, opfern
(paris, terror, victims, thoughts, relatives, attack, casualties)
Trade policy:
ttip, ceta, freihandelsabkommen, fracking, usa, abkommen, verhandlungen
(ttip, ceta, free-trade treaty, fracking, usa, treaty, negotiations)

3.2.3 Evaluating Narrative-Driven Movie Recommendations on Reddit

This article presents an evaluation of the potential of established recommender algorithms to incorporate information from narrative descriptions of users' preferences, with the aim of improving movie recommendations. Narrative descriptions of current preferences and interests are often posted on online platforms such as Reddit, by users who desire recommendations for products such as video games, movies, or board games. For example, on the subreddit r/MovieSuggestions, users post requests for movie suggestions, using natural language text to describe any number of arbitrary criteria that the recommended movies should satisfy. Other users read these requests and provide suggestions based on the description of the requester's current interests and preferences.

To evaluate the extent to which established recommender algorithms are able to incorporate information from such narrative descriptions of preferences, we first extracted relevant information from the unstructured text of the requests, employing microtasks in the *data preparation and preprocessing* stage of the machine learning process. Specifically, we employed a range of microtasks to identify movie titles as well as other relevant keywords such as genres, actors, movie settings, or other movie characteristics. Furthermore, we employed microtasks to judge the recommendation requesters' sentiment towards these movies and keywords. We then used this extracted information as well as movie metadata obtained from the Internet Movie Database to refine the computed recommendations by applying different post-filtering and re-ranking strategies.

Our evaluation of the different recommendation algorithms showed that, by using carefully configured post-filters, information extracted from narrative descriptions of users' preferences can help to greatly improve the resulting recommendations.

Evaluating Narrative-Driven Movie Recommendations on Reddit

Lukas Eberhard
Graz University of Technology
Graz, Austria
lukas.eberhard@tugraz.at

Simon Walk
Detego
Graz, Austria
s.walk@detego.com

Lisa Posch
GESIS & Graz University of Technology
Cologne, Germany
lisa.posch@gesis.org

Denis Helic
Graz University of Technology
Graz, Austria
dhelic@tugraz.at

ABSTRACT

Recommender systems have become omni-present tools that are used by a wide variety of users in everyday life tasks, such as finding products in Web stores or online movie streaming portals. However, in situations where users already have an idea of what they are looking for (e.g., *'The Lord of the Rings', but in space with a dark vibe*), most traditional recommender algorithms struggle to adequately address such a priori defined requirements. Therefore, users have built dedicated discussion boards to ask peers for suggestions, which ideally fulfill the stated requirements. In this paper, we set out to determine the utility of well-established recommender algorithms for calculating recommendations when provided with such a narrative. To that end, we first crowdsource a reference evaluation dataset from human movie suggestions. We use this dataset to evaluate the potential of five recommendation algorithms for incorporating such a narrative into their recommendations. Further, we make the dataset available for other researchers to advance the state of research in the field of narrative-driven recommendations. Finally, we use our evaluation dataset to improve not only our algorithmic recommendations, but also existing empirical recommendations of IMDb. Our findings suggest that the implemented recommender algorithms yield vastly different suggestions than humans when presented with the same a priori requirements. However, with carefully configured post-filtering techniques, we can outperform the baseline by up to 100%. This represents an important first step towards more refined algorithmic narrative-driven recommendations.

CCS CONCEPTS

• **Information systems → Recommender systems**; *Crowdsourcing*; *Personalization.*

KEYWORDS

Narrative-driven recommendations, Dataset, Crowdsourcing

ACM Reference Format:
Lukas Eberhard, Simon Walk, Lisa Posch, and Denis Helic. 2019. Evaluating Narrative-Driven Movie Recommendations on Reddit. In *24th International Conference on Intelligent User Interfaces (IUI '19), March 17–20, 2019, Marina del Rey, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3301275.3302287

1 INTRODUCTION

The practical applications of recommender systems are manifold. In general, they are tools that help users to find and discover items of interest in large collections, such as books, movies, or people. In a common collaborative filtering scenario, a recommender system makes use of a user's history and predicts new items that user is likely to read, watch, or connect to.

Problem. Often, users already have vague to specific ideas about the desired entities they want to be recommended. More precisely, users often seek recommendations that fit arbitrary criteria, such as movies that evoke certain emotions or have a surprising ending, instead of obtaining suggestions purely based on their (and other users') histories of interactions within a given system. These criteria represent the **narrative** of a recommendation request. Recommendations generated by incorporating such a narrative are referred to as narrative-driven recommendations [8] and also build the foundation for conversation-based recommendation approaches used in chat- and voice-bots. Due to the lack of automated recommender systems that can accurately calculate such recommendations, users have built various discussion boards on the Web to ask peers for suggestions. For example, as of March 2017, there were 190, 000 discussion threads with nearly 25, 000 threads containing requests with a narrative for interesting books on the social cataloging website LibraryThing[1] [8]. Also, there are several subreddits on reddit.com, where users can ask for, for example, video game, movie, or board game suggestions. Requests for movie recommendations can look as follows: *"[...] Movies with the genre 'Crime' [...] like 'Nightcrawler' and 'Prisoners' [...] And it is great if there is any form of plot twists"*[2]. The (free-form) narrative of such requests defines several different elements, such as positively or negatively associated movies

IUI '19, March 17–20, 2019, Marina del Rey, CA, USA

ACM ISBN 978-1-4503-6272-6/19/03…$15.00
https://doi.org/10.1145/3301275.3302287

[1] https://www.librarything.com

[2] https://www.reddit.com/r/MovieSuggestions/comments/3fvycr

(i.e., *Nightcrawler*, *Prisoners*), preferred as well as unwanted genres (i.e., *Crime*), and specific keywords that define desired or undesired attributes/keywords of the movie (i.e., *plot twists*) [8].

Approach. In this paper, we systematically analyze the suitability of five standard recommender algorithms for supporting such a narrative in recommender systems. For our evaluation, we compare human suggestions for requests that provide a narrative with purely algorithmic recommendations.

To that end, we first compile an evaluation dataset by collecting and parsing narrative requirements from users of the subreddit r/MovieSuggestions[3]. We extract requirements from the unstructured text of submissions and comments with the help of crowdworkers and make our dataset available online[4] for future research. Next, we implement a recommender framework based on ratings, reviews and textual information of movies available on the Internet Movie Database[5] (IMDb). We calculate recommendations using the following five algorithms for our analysis: item-based collaborative filtering (CF), matrix factorization (MF), a content-based filtering approach based on TF-IDF similarities (TF-IDF), document-level embeddings (Doc2Vec), and a network-based approach (NW). In addition, we extract movie suggestions generated by IMDb, which we use as an empirical baseline (IMDb baseline). We apply post-filtering and re-ranking strategies using metadata from IMDb to refine the computed recommendations. Finally, we evaluate the five recommender approaches by measuring the overlap between their recommendations and the suggestions from users in our evaluation dataset from reddit. Our initial results suggest that traditional recommender algorithms exhibit great potential for improvement when presented with a narrative, as they lack the proper means to include a priori specified requirements in the recommendation process. Further, we demonstrate that we can improve all recommendation approaches (including existing empirical IMDb recommendations) by applying post-filtering and re-ranking strategies using metadata available in the narrative of the initial requests on reddit.

Contributions. With our analyses, we make the following contributions. First, we publish a reference dataset, which enables researchers to conduct independent analyses, advancing the state of research in the context of narrative-driven recommendations. Second, we evaluate the performance of five well-studied recommender approaches on our reddit evaluation dataset, containing a total of 1, 480 recommendation requests that provide a narrative. Third, we demonstrate how to improve narrative-driven recommendations by introducing post-filtering and re-ranking techniques and analyze their importance for each of our five implemented recommendation approaches.

2 RELATED WORK

Traditional Recommender Systems. There exists a vast variety of studies about recommender systems and algorithms (e.g., [3–5, 8, 10, 11, 14, 15, 18, 20–22, 27, 34]). However, we still only have limited insights into the quality and suitability of traditional recommender algorithms for calculating narrative-driven recommendations. Typically, traditional research in recommender systems focuses on algorithmic advantages in common scenarios, such as applying users' histories and profiles to compute recommendations [11, 14, 21, 27].

Context-Aware Recommender Systems. To compute recommendations that are well suited to the current needs of a user, context-aware recommender systems use contextual information, such as the time of the day or the current location or interests of the user, besides user profiles and histories [15]. In a context-driven environment, Adomavicius et al. [4] introduced REQUEST, which is a query language for customizing recommendations based on users' personalized recommendation needs. Hariri et al. [15] proposed a query-driven context-aware recommender system that considers user profiles, item representations, and contextual information, such as interests or needs of a user in a specific situation.

A context-aware support vector machine for application in a context-dependent recommender system was proposed by Oku et al. [24]. The authors found that for information recommendation it is important to consider the situations or conditions which influence the users' decisions (e.g., time of day, weather, physical condition).

In the study of Adomavicius et al. [1], the authors presented a multidimensional recommendation model that is based on additional contextual information, such as profiles and aggregation hierarchies. They evaluated their approach on a movie recommender by exploiting contextual information, such as when a movie was seen, where, and with whom. They empirically demonstrated that this contextual information can improve the recommendations.

Basu et al. [5] conducted a study on IMDb data, in which they proposed a recommender approach that exploits both user ratings and content information using collaborative, content, and hybrid features. Lamprecht et al. [18] analyzed how IMDb recommendation networks support alternative information retrieval strategies, such as browsing. The authors showed that current recommendation networks are poorly navigable and require further improvements. This shows potential for providing context-aware recommender systems that involve the current needs of a user without the need of clicking through poorly navigable recommendation networks until finding a more or less fitting movie.

Adomavicius and Tuzhilin [3] argued that relevant contextual information is important when providing recommendations. Such contextual information can be obtained explicitly (i.e., users provide additional information) or implicitly (i.e., system implies the context automatically from the given requirements). To that end, the authors introduced pre- and post-filtering techniques for capturing relevant context during the recommendation process. They used these methods for selecting a relevant set of data and for filtering out irrelevant recommendations or adjusting the ranking of the obtained recommendation list based on a given context. They discussed the notion of context and how it can be modeled, and conducted an empirical analysis using movie data regarding only the combination of several pre-filters. In this paper, we follow up on their ideas.

In contrast to the study of Panniello et al. [25] that constitutes a first step towards the comparison of pre- and post-filtering using

[3] https://www.reddit.com/r/MovieSuggestions
[4] http://www.rbz.io/datasets
[5] https://www.imdb.com

just one contextual variable for each applied dataset, we introduce and combine several post-filters and evaluate their utility in the context of narrative-driven movie recommendations.

Narrative-Driven Recommender Systems. Bogers and Koolen [8] presented a specific context-aware recommendation scenario called narrative-driven recommendation. In such a scenario recommendations are computed based on past transactions of users, and a narrative description of the current needs and interests of users. Narrative-driven recommendations are related to conversational-based recommender systems, where users ask for suggestions in a community and other users then come up with suggestions and possible explanations for their choices [10, 20, 22].

Bogers [7] analyzed the movie discussion threads from the IMDb message boards that contain requests for movies to watch. The author found that content (e.g., movie description), different types of metadata (e.g., genre, language, release year), and searching for a movie by describing its content (e.g., in cases where users forgot the movie title) are important for movie selection practices.

In contrast to previous work, we present the first in-depth analysis and evaluation of recommender algorithms to support narratives for the computation of recommendations.

3 REDDIT NARRATIVES EVALUATION DATASET

On r/MovieSuggestions, users ask other users for movie suggestions by describing, in natural language, what they are looking for. For example, typical posts include questions such as *“[...] Really dark, slow paced movies with minimal story, but incredible atmosphere, kinda like ‘Drive’ (2011), ‘The Rover’ (2014), or ‘No Country for Old Men’ (2007)? [...]”*[6]. The narrative of this example includes references to three “positively associated” movies (i.e., *Drive*, *The Rover*, *No Country for Old Men*) and several keywords that define the gist of the plot (i.e., *incredible atmosphere, dark, slow paced, minimal story*). As these requests are written in free-form text, the amount of information that can be leveraged for calculating recommendations varies. For example, users sometimes include detailed lists and descriptions of movies that they previously did (or did not) enjoy in their requests. Other times, only a single movie is referenced. Further, users frequently provide keywords in the narrative, which should apply to the suggestions (e.g., *“[...] Movies that will make me want to cry [...] like ‘Extremely Loud and Incredibly Close’ ”*[7] with the keyword *cry* and one desired movie, or *“[...] Movies that take place primarily in one room or building. [...] Examples: Exam, Circle, Hateful Eight, Die Hard [...]”*[8] including the keywords *one room or building* and some desired movies). Other users then suggest appropriate movies by writing comments to the original post. Note that recommendations on r/MovieSuggestions are usually generated only considering the information provided in each submission, ignoring previous interactions or requests of users, limiting the amount of available information (see Table 1 for a more detailed characterization of our dataset).

Requests with a Narrative. To compile a dataset suitable for the evaluation of narrative-driven recommendations, we extracted all submissions from r/MovieSuggestions that (i) were posted between August 14, 2011 and August 1, 2017[9], (ii) had received at least ten comments, and (iii) had a score (i.e., the sum of up- and downvotes) greater than zero (3,640 of 23,484 submissions after filtering). Additionally, we extracted all comments to these submissions that had a score greater than zero, which we used as indicator for good recommendations (24,851 of 201,298 comments after filtering). For the compilation of the dataset, we asked crowdworkers to match the movies, genres, actors and other keywords mentioned in the reddit narratives to their corresponding entries on IMDb. The IMDb website provides a wide variety of information about movies and TV shows, such as genres, descriptions, trailers, plot summaries, as well as details about the cast, producers, and writers. In February 2017 the publicly available dataset[10] included information about 4.1 million titles and 7.7 million people.

Crowdsourcing Requests and Suggestions. To obtain a structured set of user requests and suggestions, we asked crowdworkers to annotate the unstructured text of the previously extracted submissions and comments from r/MovieSuggestions after filtering (see Table 1 for more details). To that end, we designed four micro tasks

Table 1: *Reddit Evaluation Dataset Characteristics.* This table lists the statistics of our reference dataset, which we compiled using data from r/MovieSuggestions and crowdworkers on Crowdflower to extract structured data from the unstructured text of the submissions and comments.

#Submissions	1,480
Average Submission Score	11.78
#Movies in Submissions	5,521
#Unique Movies in Submissions	1,908
#Submissions with Desired Movies	1,480
#Submissions with Undesired Movies	75
#Keywords in Submissions	4,492
#Unique Keywords in Submissions	1,878
#Submissions with Desired Keywords	1,198
#Submissions with Undesired Keywords	153
#Genres in Submissions	762
#Unique Genres in Submissions	26
#Submissions with Desired Genres	491
#Submissions with Undesired Genres	61
#Actors in Submissions	100
#Unique Actors in Submissions	79
#Submissions with Desired Actors	75
#Submissions with Undesired Actors	6
#Comments	21,032
Average Comment Score	2.88
#Movie Suggestions in Comments	43,402
#Unique Movie Suggestions in Comments	6,071
Average #Movie Suggestions per Submission	29.33
Average #Movie Suggestions per Comment	2.48

[6] https://www.reddit.com/r/MovieSuggestions/comments/3kjrus

[7] https://www.reddit.com/r/MovieSuggestions/comments/11ycep

[8] https://www.reddit.com/r/MovieSuggestions/comments/4va9p8

[9] The dump is available at https://files.pushshift.io/reddit [6]

[10] https://www.imdb.com/interfaces

on the crowdsourcing platform CrowdFlower (now Figure Eight).[11] First, in the SUBMISSIONS task, we asked crowdworkers to identify all movie titles in each submission. Second, in the SENTIMENT task we asked crowdworkers to specify the sentiment of the user with respect to a movie mentioned in a submission (i.e., positive or negative association to the requested suggestions). We defined *positively associated* movies as movies that users liked or where they stated that they were looking for movies similar to these. Analogously, we defined *negatively associated* movies as movies that users disliked or where they stated that they were not looking for similar movies. Third, in the KEYWORDS task we asked crowdworkers to identify additional information about the user's preferences in each submission's text (i.e., keywords). To extract these keywords, we provided the crowdworkers with a list of keyword types containing, for example, genres, movie settings, and events.[12] We asked the crowdworkers to identify *positively associated keywords* (i.e., keywords which should apply to the recommendations) and *negatively associated keywords* (i.e., keywords which should not apply to the recommendations). Finally, in the COMMENTS task, crowdworkers identified all movie titles in the comments to each submission.

A minimum of three separate crowdworkers worked on each submission in the SUBMISSIONS task. Where there was high disagreement among the workers, we requested judgements from two additional workers. Three workers worked on each movie in the SENTIMENT task and each comment in the COMMENTS task. In the KEYWORDS task, five distinct workers extracted keywords from each submission. We ensured the quality of the crowdworkers' output by requiring an entry-quiz for each task. Additionally, we continuously assessed workers via test questions.

Post-Processing. To obtain a well-curated dataset for the training and evaluation of narrative-driven recommendations, we carried out several manual and semi-automatic post-processing steps.

First, we manually reviewed all submissions from the SUBMISSIONS task and all comments from the COMMENTS task that did not have the crowdworkers' full agreement on movie titles. The crowdworkers fully agreed on the movie titles in 1, 205 submissions and 16, 893 comments, and they disagreed on titles in 457 submissions and 7, 958 comments, which we then manually reviewed. During this step, we also removed submissions and comments without movie titles.

Second, we aggregated the answers from the SENTIMENT and KEYWORDS tasks. In the SENTIMENT task we applied a majority vote whereas in the KEYWORDS task we first split the keyword strings provided by the workers into single keywords. Then, we retained all keywords identified by at least two out of the five workers.

Third, we automatically and unambiguously matched 1, 298 movie titles from the SUBMISSIONS and 5, 695 movie titles from the COMMENTS task to movie titles from IMDb. We then manually reviewed all movie titles that could not be automatically mapped to IMDb. In cases where more than one (or no) movie existed with the exact same movie title, we matched the movie using contextual information of the submission and the comments. In cases where we did not have sufficient information to unambiguously map movies, we removed them from our reference dataset.

Fourth, we automatically identified all common movie genres and actors in the keywords by matching them to the 25 genres and 294, 533 actors available in our IMDb data.

Finally, we removed all movies from the submissions and comments that are not present in our IMDb data. Further, we removed submissions that did not contain any positively associated movie and that did not receive at least ten unique movie suggestions in the comments. After the last preprocessing step, our reference dataset[13] consists of 1, 480 movie-recommendation requests and 43, 402 corresponding suggestions, as noted in Table 1.

4 EXPERIMENTAL SETUP

Our recommendation framework (see Figure 1) (i) uses one or more movies as input data, (ii) implements five different recommender algorithms to compute a candidate set of recommendations, and (iii) applies several post-filtering and re-ranking strategies, based on metadata from IMDb to calculate a final list of (top ten) recommendations.

To assess the importance of narratives for the calculation of recommendations we further calculate an alternative final recommendation list by applying the structured input (in the form of e.g., actors and keywords) from a given reddit narrative in the post-filtering and re-ranking step.

Finally, we evaluate both lists by comparing them to human suggestions from our reddit evaluation dataset (see Section 3).

Hyperparameter Optimization. To analyze if and to what extent traditional recommender approaches can support narratives we aim at making as few assumptions as possible and take a data-driven approach. Thus, we conduct an extensive cross-validation over various configurations of the parameters of the algorithms (see framework components highlighted in orange in Figure 1). Specifically, we optimize (i) hyperparameters for the algorithms, such as similarity measures or regularization parameters, (ii) the lengths of the initial and the final recommendation lists, and (iii) hyperparameters of the post-filtering and re-ranking mechanisms, such as overlap measures or functional forms for various scores. We discuss the optimal parameter configurations that we obtain along with introducing a given framework component.

IMDb Movies & Ratings. To implement the recommender algorithms we use data from IMDb. Note that training of recommender algorithms directly on our reddit evaluation dataset is not viable due to the sparsity of data. We leave this option open for future work when more data is available.

In addition to the publicly available IMDb dataset, we collect user reviews and individual ratings for all movies on IMDb. For our experiments, we only consider movies and discard all other types available on IMDb, such as TV series or single TV episodes. To minimize noise and to allow for fair comparisons between the different approaches, we only *keep movies* that have (i) more than 1, 000 user ratings, (ii) at least one user review, (iii) a movie description, and

[11] https://www.figure-eight.com

[12] The full list of keyword types included *genres, actors, movie directors, movie characters, movie producers, movie production companies, events or special occasions, movie settings,* and *other movie characteristics.*

[13] Note that on our website http://www.rbz.io/datasets, we also provide necessary information about the mapping of genres and actors, and an extended version of our dataset without thresholds for the number of suggestions or the number of positively mentioned movies.

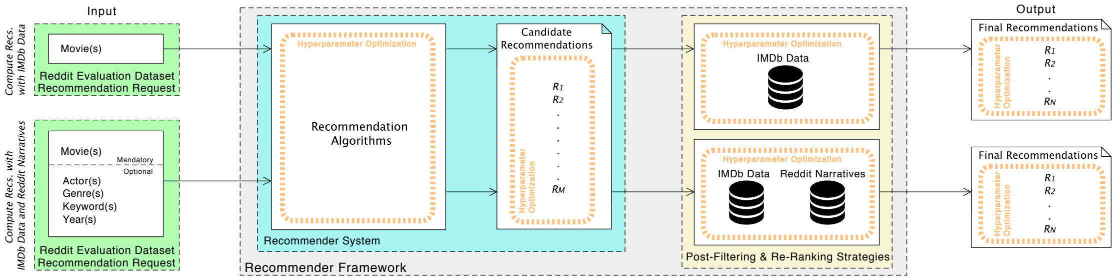

Input
Compute Recs. with IMDb Data
Movie(s)
Reddit Evaluation Dataset Recommendation Request
Compute Recs. with IMDb Data and Reddit Narratives
Movie(s)
Mandatory
Optional
Actor(s)
Genre(s)
Keyword(s)
Year(s)
Reddit Evaluation Dataset Recommendation Request
Hyperparameter Optimization
Recommendation Algorithms
Candidate Recommendations
R1
R2
RM
Hyperparameter Optimization
Recommender System
Recommender Framework
Hyperparameter Optimization
IMDb Data
Hyperparameter Optimization
IMDb Data
Reddit Narratives
Post-Filtering & Re-Ranking Strategies
Output
Final Recommendations
Hyperparameter Optimization
R1
R2
RN
Final Recommendations
Hyperparameter Optimization
R1
R2
RN

Figure 1: *Experimental Setup.* The recommender framework accepts several input parameters (see *Input*), extracted from the narrative of a recommendation request (e.g., reddit submissions). We distinguish between requests that only provide information about desired movies (see *Compute Recs. with IMDb Data*) and requests that include more detailed information from their narratives (see *Compute Recs. with IMDb Data and reddit Narratives*). The input parameters are then fed into the implemented recommender algorithms (see *Recommender System*), which calculate a first list of candidate recommendations. We then apply post-filters (see *Post-Filtering & Re-Ranking Strategies*) based on IMDb Metadata, or IMDb Metadata and reddit narratives, to provide a re-ranked list of recommendations (see *Output*), which better reflects the requirements defined in the narrative of the recommendation request. For all parts that are highlighted in orange (see *Hyperparameter Optimization*), we conduct an extensive grid search over relevant parameter configurations to find the optimal parameter settings.

(iv) at least one person in the cast. In contrast, we do not *remove users* with small numbers of ratings, as this preprocessing step does not improve our results. We obtain the rating thresholds for movies (1, 000) and users (no limit) via grid search.[14] For more details see Table 2. Further, we compute centered ratings [12, 29] by removing user and item bias which improves the overall performance of all implemented recommender approaches.

4.1 Recommender Strategies

We generate recommendations by computing similarities between an input movie and all other movies available in our IMDb dataset. Each recommender algorithm determines how and with which data we calculate similarity. As similarity measures we use cosine similarity and an inverse of Euclidean distance and select the best performing measure via cross-validation. In cases where we have more than one input movie we aggregate similarity values. Hence, for each movie in our IMDb data, we add all similarities for all positively associated input movies. Our cross-validation yields better results when we do not subtract negative input movies for the aggregation of similarity values. We call the aggregated similarities *algorithmic score.* Thus, the output of each approach is a ranked list of candidate movies with their corresponding algorithmic scores. We conduct experiments with the following five approaches:

Item-Based Collaborative Filtering. This approach finds similar movies to the movies that a user liked [33]. Thus, we use the IMDb user-ratings vectors of two movies to compute their similarity [33]. The best performing similarity measure for this approach is cosine similarity.

Matrix Factorization. This approach is a well-established method that approximates a ratings matrix with the product of two matrices, one connecting users to factors representing their preferences, and another connecting movies to factors representing their properties [17, 26, 30, 31]. In this paper, we factorize the IMDb user-ratings matrix in a standard manner by minimizing a regularized squared error with a stochastic gradient descent [13]. We then use cosine

Table 2: *IMDb Dataset Characteristics.* This table describes the features of the dataset that we used for computing the recommendations of our implemented recommender algorithms.

#Movies	11, 578
#Ratings	144, 021, 151
Average #Ratings per Movie	≈ 12, 439.21
#Users with Ratings	1, 144, 136
Average #Ratings per User	≈ 125.88
#Reviews	1, 880, 837
Average #Reviews per Movie	≈ 162.45
#Users with Reviews	598, 247
Average #Reviews per User	≈ 3.14
#Credits	667, 279
#People in Cast & Crew	322, 881
#Actors	294, 533
Average #Actors per Movie	≈ 25.44
Average #Movies per Actor	≈ 2.27
#Genres	32, 767
#Unique Genres	25
Average #Genres per Movie	≈ 2.83
#Plot Keywords	1, 124, 510
#Unique Plot Keywords	89, 003
Average #Plot Keywords per Movie	≈ 97.12

[14] We perform the grid search over 0 to 10, 000 movie ratings in increments of 500, as well as 0 to 500 user ratings with increments of 10.

similarity (determined via hyperparameter optimization) to compute similarity between the obtained movie factors.[15]

Content-Based Filtering with TF-IDF. We use this approach to find similar movies by calculating similarity between movies using their descriptions and user reviews [2]. Hence, we compute the term frequency–inverse document frequency score [32] of terms in the description and user reviews for each movie. To compute the similarity between movies we use normalized TF-IDF vectors and the reciprocal of Euclidean distance (determined via hyperparameter optimization). We receive the best results with unigrams and bigrams, no cut-off threshold for less frequent terms, and with a maximum of 500 features for the TF-IDF vectors.[16]

Document-Level Embeddings with doc2vec. Similar to the TF-IDF approach, we use movie descriptions and reviews as basis for this approach. doc2vec was first proposed by Le and Mikolov [19] and is an enhancement of word2vec [23], extending the learning of embeddings from words to documents. We use doc2vec to generate a document vector for each movie and use these vectors to compute similarities between movies. We obtain the best results with a feature vector dimensionality of 500 and cosine similarity.[17]

Network-Based Recommendations. We use this approach to find movies with similar casts and crews by creating a bipartite graph between movies and people involved in those movies. Specifically, we connect each movie to all cast and crew members including actors, cinematographers, composers, costume designers, directors, editors, producers, production designers, special effect companies, and writers. We calculate similarity between movies by counting common neighbors in the bipartite graph [16].

IMDb Baseline. We collect all movie suggestions on IMDb[18] for each movie in our dataset to determine if and to what extent existing (empirical) recommender systems are suitable to address a narrative. IMDb provides a maximum of twelve recommendations per movie. We use these recommendations for all (desired) input movies in the narrative of each submission. Note that IMDb does not provide any ranks or numerical values quantifying the quality of each recommendation.

4.2 Post-Filtering & Re-Ranking

We further refine the algorithmic recommendations by defining several post-filtering approaches, which allow us to include (i) additional metadata from IMDb, and (ii) optionally reddit narratives in our recommendations. Again, for evaluation of various post-filters we pursue a data-driven approach and conduct extensive cross-validations over multiple configurations. This allows us to evaluate the importance of the individual post-filters as well as the interactions between different post-filters.

Specifically, with our post-filtering techniques we modify the calculated recommendation list by (i) removing irrelevant recommendations for a given movie, and (ii) re-ranking the obtained list. In general, the more properties (e.g., genres, keywords, actors) the candidate movies have in common with a given input movie, the higher they get ranked. For example, we compute the overlap of genres of all input movies and a candidate movie. With all scores calculated we re-rank the candidate lists by combining algorithmic scores of each candidate recommendation with the corresponding post-filtering scores to compile a final recommendation list. We evaluate the resulting (final) list by comparing it to human suggestions from our reddit evaluation dataset. When limiting our final recommendation list to a total of ten movies to be displayed, we achieve the best results with 500 candidate recommendations.[19]

Post-Filtering & Re-Ranking with IMDb Data. With IMDb metadata we re-rank candidate recommendations with the following scores:

IMDb Popularity & Rating Score. Following the intuition that users are generally more interested in higher and more frequently rated movies, we introduce this score which combines the average IMDb rating (*rating score*) of a candidate movie and the number of ratings received on IMDb (*popularity score*). We experiment with various functional forms for the computation of both the average rating and the number of ratings. Specifically, we calculate logarithmic, square root, quadratic, and cubic scaling and achieve the best results with the following functional form: $\log_2(R_i)\overline{r}_i$, where $\overline{r}_i$ is the average rating, and R_i is the number of ratings of movie i.

IMDb Genre Score. Here, we follow the intuition that users prefer movies of similar genres to the specified movies and calculate the IMDb genre score for each candidate movie. As part of our hyperparameter optimization, we compare several overlap measures, including Jaccard's coefficient, cosine similarity, Sørensen-Dice coefficient, and simple matching coefficient. We achieve the best results with similar scaling and normalizing of the overlap between the genres of the candidate movie and individual positively associated movies from the request so that $S_{\text{iGenre}}(i) = \sum_{j \in I_{\text{pMovie}}}(|G_i \cap G_j|^2/(|G_i||G_j|))$, where I_{pMovie} is the set of positively associated input movies. The inclusion of negatively associated input movies does not improve our results.

IMDb Year Score. We assume that users want to watch movies from similar time periods unless explicitly stated otherwise. Thus, we introduce the IMDb year score, where candidate movies released closer in time to the input movies receive higher scores. We set this score to 1 for a candidate movie with the smallest difference in release year to one of the input movies. We then linearly scale the year score until we reach 0 for a given maximal difference in release years. We obtain the best results with a release year normalization of 50 years.[20]

IMDb Keyword Score. For our recommender framework, keywords are words or phrases that represent a very specific attribute of a movie. For the IMDb keyword score, we use the plot keywords from IMDb and compute the overlap of all plot keywords of a candidate movie and the plot keywords of the input movies. Following a grid

[15] We have tested different numbers of factors ranging from 100 to 1, 000 in steps of 100, learning rates between 0 and 0.1 in steps of 0.001, and regularization parameters from 0 to 0.1 in steps of 0.01. We obtain the best results for MF with 500 factors, a learning rate of 0.002, and 0.02 as regularization parameter.

[16] To obtain this configuration we conducted a grid search experiment over different n-grams [9] (i.e., n = 1, 2, 3), several cut-off values for terms with a low document frequency from 0 to 0.1 in increments of 0.001, and different numbers of TF-IDF features ranging from 0 to 1, 000 in steps of 100.

[17] To obtain this configuration we conducted a grid search experiment with different similarity measures, and different feature vector sizes ranging from 0 to 1, 000 in steps of 100.

[18] For an example see "More Like This" on https://www.imdb.com/title/tt0076759

[19] We determine the length for the candidate list with a grid search over the range from 100 to 1, 000 movies in steps of 100.

[20] Identified via grid search over the year-range from 20 to 100 in steps of 10.

search approach we determine Jaccard's coefficient as the most suitable overlap measure while ignoring plot keywords of negative input movies.

IMDb Predecessor & Successor Filters. We assume that users do not want to receive a list of predecessors or successors of the specified input movies as they are likely familiar with the whole series. Hence, we remove predecessor and successor movies from our recommendation lists. For example, if users ask for movies similar to *The Hunger Games: Catching Fire* we remove *The Hunger Games* and *The Hunger Games: Mockingjay - Part 1 & 2* from our recommendation list.

Combining Scores. To compute the final score for each candidate movie we first normalize all computed scores by their highest values, so that (for each score individually) the movie with the highest score receives the value 1. Second, as post-filters are not equally important across our approaches, we multiply the scores with weights, reflecting their influence for the re-ranking of the recommendation lists. We conduct a grid search experiment over all combinations of weights between 0.0 and 1.0 in steps of 0.2, and select the setup that yields the best results in our experiments. Finally, we sum up all weighted scores to obtain the final score for each movie.

Post-Filtering & Re-Ranking with Reddit Narratives. For the final step of our evaluation we incorporate metadata, available in the narrative of the initial reddit submission, into our recommendations using additional post-filters. Specifically, we use keywords, genres, actors, and years given in the narrative of the movie suggestion requests in our reddit evaluation dataset. Note that we can calculate post-filtering scores from reddit narratives only if users explicitly provided positively/negatively associated attributes or keywords (e.g., actors or genres) in a recommendation request (see Table 1). With all scores calculated we re-rank the candidate lists (500 candidates) again by combining all IMDb post-filtering scores of each candidate recommendation with the corresponding narrative-based post-filtering scores to compile a final recommendation list (ten recommendations). Again, we evaluate the resulting (final) list by comparing it to human suggestions from our reddit evaluation dataset. To that end, we define and compute the following narrative-based post-filtering scores and evaluate their importance by conducting a grid search experiment:

Narrative-Based Genre Score. If genres are stated in the narrative of a request, we use them to calculate the narrative-based genre score for each candidate movie. We ran the same grid search experiment as we did for the *IMDb Genre Score* and determined that the same overlap metric yields the best results. In contrast to the *IMDb Genre Score*, we remove movies with undesired genres from our recommendation list.

Narrative-Based Year Filter. If users explicitly state year thresholds, we re-rank the recommendation list so that movies outside this range are moved to the end of the list.

Narrative-Based Keyword Score. We exploit keywords in a specific request (e.g., "surprising plot twist") to introduce the narrative-based keyword score. With this score we measure how well the description and the user reviews of a candidate movie reflect the keywords stated in a narrative. We find that counting the incidences of explicitly stated keywords in the description and all user reviews of the respective candidate movies yields the best results by conducting a grid search experiment. We aggregate the incidences for positive input keywords and subtract them for negative ones. Finally, we compute the narrative-based keyword score by normalizing over the number of words in the used texts.

Narrative-Based Actor Filter. To reflect the requirement of only recommending movies with specific actors, we introduce the actor filter. We re-rank the list of movie recommendations by counting how many of the positively stated actors appear in the respective movies. Further, we remove all movies with actors that users explicitly specified as undesired.

Combining Scores. To combine all narrative-based post-filtering scores we use the same method as for the IMDb post-filtering scores.

4.3 Evaluation

We evaluate the implemented approaches on our reddit evaluation dataset. Specifically, we use the narrative from each submission to calculate movie recommendations and count the overlap between the movie suggestions of the reddit community, extracted from the replies to the corresponding submission (see Section 3) and our algorithmic movie recommendations. We calculate precision, recall, F1 score, normalized discounted cumulative gain (nDCG), and mean average precision (MAP) [28, 35]. First, we chronologically split our reddit evaluation dataset into a validation (80%) and a test (20%) set (see Table 3). Second, we train our approaches on the IMDb data and use the reddit data from the validation set to conduct all grid search experiments for optimizing hyperparameters for the recommender framework. Finally, we evaluate the performance of the implemented approaches on the test set. We limit our final recommendation lists to ten movies.[21] To allow for a fair comparison we also limit the number of recommendations for our IMDb baseline to ten movies (picked at random, as recommendations are not ranked). First, we evaluate the standard algorithms with post-filters and scores calculated by using IMDb metadata. Second, we measure the performance improvements with the narrative-based post-filters and scores.

5 RESULTS & DISCUSSION

5.1 Post-Filtering & Re-Ranking with IMDb Data

Figure 2 depicts the results of the evaluation of our implemented algorithms for calculating recommendations for a given narrative using our reddit evaluation dataset. The transparent bars represent

Table 3: *Evaluation Protocol.* Basic statistics of the validation set and the test set.

	#Submissions	Timeframe
Validation Set	1, 184	08-2011 – 11-2016
Test Set	296	11-2016 – 07-2017
Overall	1, 480	08-2011 – 07-2017

[21] This means that recall@10 and F1 score@10 have a mean upper limit of 0.34 and 0.51 respectively, as the average number of movie suggestions from the community per submission is 29.22 in the test set.

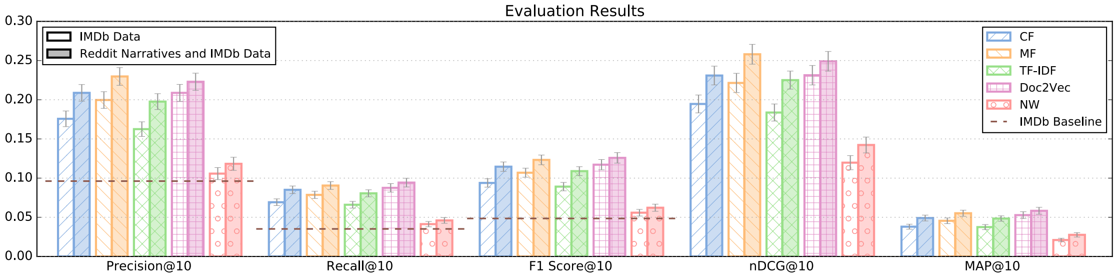


Figure 2: *Results.* This figure depicts the results of our evaluation, comparing our recommendations to the ones of the reddit community in our reddit evaluation dataset. We list the different evaluation metrics on the x-axis, with the corresponding evaluation metric values on the y-axis. The performances of the recommender algorithms with IMDb post-filters are represented by the transparent bars, while the filled bars depict the results for the approaches with additional narrative-based post-filters using reddit data. The grey error bars show the standard deviation of the evaluation metric over all submissions in the test set. All of our approaches outperform the IMDb baseline (dashed horizontal line). We can further improve the results by adding narrative-based post-filters, where Doc2Vec outperforms all other approaches with F1 scores more than twice as good as the IMDb baseline.

the means of the evaluation metrics over all submissions in the test set for a given approach using only IMDb-based post-filters, with the error bars showing the standard error. All of our analyzed approaches, while only relying on IMDb-based post-filters, manage to outperform the IMDb baseline (cf. horizontal dashed line in Figure 2). Doc2Vec performs best in all evaluation metrics with an F1 score of 0.117, which is more than twice as good as the IMDb baseline, followed by MF with 0.107, CF with 0.094 and TF-IDF with 0.089, and NW, which performs consistently worst with an F1 score of 0.056, while still outperforming the IMDb baseline.

One possible reason for the moderate performance of NW might be that this approach is fundamentally based on the assumption that users want to see other movies with a similar cast. This inherent restriction appears to impair our results when incorporating the narratives provided by users. However, more research is warranted to further investigate this hypothesis, which we leave open for addressing in future work. MF and CF perform roughly twice as good as NW, possibly due to the larger amount of considered data. They are both based on user ratings and follow similar intuitions (i.e., both approaches favor frequently and highly rated movies), which could explain the similarity in the obtained results. TF-IDF, which is based on the text of movie descriptions and user reviews, performs similar to CF. Doc2Vec performs best of all approaches using the same data, which we attribute to the underlying mechanisms of the approach. Compared to TF-IDF vectors, word embeddings better incorporate latent factors in textual representations, leading to better similarity calculations and, therefore, better recommendations.

Importance of Post-Filters. We present the best-performing IMDb-based post-filter configuration for each approach by depicting the normalized score weight for each post-filter in Figure 3 (obtained by cross-validation), where a higher score signals higher importance of a given post-filter.

In case of CF, we obtain the best-performing configuration with a weight of 0.8 for the algorithmic score, a weight of 0.0 for IMDb popularity and rating influence, and relatively low weights of 0.4 for IMDb genre, keyword and year scores. For MF a higher algorithmic score weight (1.0) and high popularity and rating influence of 0.8 work best, while the year score is completely neglected and the IMDb genre and keyword scores are set to 0.2 and 0.4, respectively. The content-based approaches (TF-IDF and Doc2Vec) exhibit similar best-performing configurations with a 1.0 weight for the algorithmic scores, a 0.6 weight for the IMDb popularity and rating scores and a 0.2 weight for the IMDb year scores. The weights for the IMDb genre and keyword scores range between 0.0 to 0.4. In contrast, NW mainly relies on keywords and popularity and rating influence with weights of 0.6 for the algorithmic score, 1.0 for the IMDb popularity and rating score and 0.8 for the IMDb keyword score. Similar to most other approaches, the influence of IMDb genres and years is quite low.

Findings. Our results reveal that for narrative-driven recommendation scenarios traditional recommender algorithms exhibit only minimal overlaps with human suggestions. Specifically, the algorithmic recommendations using post-filtering with IMDb metadata are computed by calculating similarities between the input movies and the movies from our dataset, while the narrative from reddit is neglected. However, additional information provided by users within their submissions appears to be crucial for the selection of appropriate movie suggestions. Users on reddit parse and consider this information, discerning their recommendations from algorithmic ones.

5.2 Post-Filtering & Re-Ranking with Reddit Narratives

In Figure 2 we also show the results of our experiments with post-filtering and re-ranking of the recommendations using the information from reddit narratives. Due to the fact that we now include narratives we can observe substantial improvements of our results when adding—and carefully configuring—post-filtering techniques (cf. transparent versus color-filled bars in Figure 2). Although not

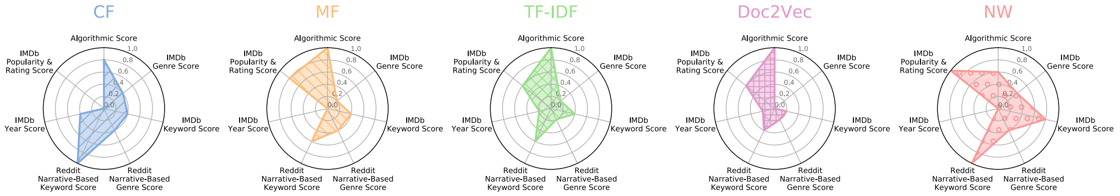


Figure 3: *Score Weights.* Each figure visualizes the score weight configuration of one approach. The algorithmic score and the IMDb popularity and rating scores are important characteristics across most of our approaches. Using narrative-based post-filters, the most important property are the keywords with weights up to 1.0. This also indicates that keywords are important for calculating narrative-driven recommendations.

exhausting the potential for improvement, we raise F1 scores of our approaches to be more than twice as high as the IMDb baseline, except for NW. Again, we achieve the best results using Doc2Vec with an F1 score of 0.126, closely followed by MF with 0.123, CF with 0.115 and TF-IDF with 0.109.

Importance of Post-Filters. Although the inclusion of the narrative information improves the recommendations, this additional information needs to be properly configured and strongly depends on the underlying algorithm. For all approaches, the best-performing configuration exhibits higher score weights for keywords extracted from the reddit narratives than for genres. For CF and NW, the narrative-based keyword score is very important, with configuration weights of 1.0, while it is 0.6 for MF and TF-IDF and 0.2 for Doc2Vec. For the narrative-based genre score CF, MF and NW have the same weights of 0.4, while the content-based approaches (TF-IDF and Doc2Vec) exhibit lower score weights of 0.2.

Findings. We find that carefully weighing the different post-filters, particularly in combination with the algorithmic, popularity and rating score, is important to maximize the benefit of the additional information contained in a given narrative.

Further, we find that for all approaches the most important narrative-based post-filter is the keyword score. From this result, we conclude that narrative recommendation requirements, provided in the form of keywords (i.e., the gist of a given text, such as short aspects of the story of a movie), are integral for achieving the best recommendations in our setup. We hypothesize that these keywords provide our post-filters with important information, that specifically helps to filter noise (i.e., unwanted movies) and steer our results towards more fitting movies. However, more research is warranted not only to confirm our hypothesis, but also to determine if additional post-filter or re-ranking strategies exist, for example, based on analyzing characteristics of recommendation requests, which could help to further improve our results.

Besides the narrative-based keyword score, the algorithmic and popularity and rating scores are also important for most of our approaches. This finding also strengthens our intuition that the configuration of algorithmic scores and post-filters is important for the computation of narrative-driven recommendations, and that it is not sufficient to simply apply filters on a given pool of existing recommendations as valuable information is lost and neglected in that process.

Except for NW, the influence of the IMDb genre and keyword scores are similarly low across all approaches. The least important score is the IMDb year score with weights ranging from 0.0 to 0.4. In fact, after manually inspecting our dataset, it appears that movies suggested by humans are more frequently from different years (even decades) than the movies mentioned in the recommendation requests (i.e., reddit submissions).

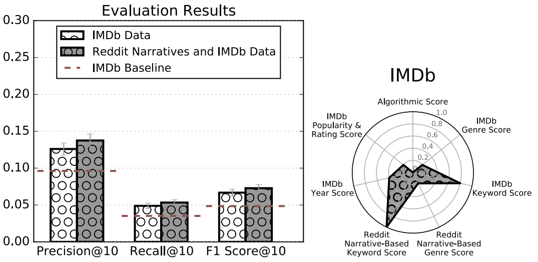
5.3 Applying Post-Filters on Empirical Recommendations

In addition to the datasets presented in this paper, we conduct another experiment to see if our post-filtering strategies can also improve our baseline IMDb recommendations. To that end, we apply all our post-filters on the IMDb baseline. We deploy the same evaluation setup as for our other previous experiments. First, we conduct a grid-search experiment to achieve the best-performing post-filter weights combination. Second, we apply all IMDb post-filters on the IMDb recommendations list and use the top ten recommendations for evaluation. The results, represented by the transparent bars in Figure 4a, reveal that additional IMDb metadata can be used to improve the resulting recommendations. Finally, we add post-filters with metadata from reddit narratives to the IMDb recommendations and further improve our results (see filled bars in Figure 4a), showing that it is possible to refine and improve recommendation algorithms to better support a given narrative using the post-filters presented in this paper.

The most important post-filters for this approach are the keyword scores from the IMDb data as well as from the reddit narratives (see Figure 4b). This further strengthens our finding that keywords provided in narratives are an important factor when re-ranking recommendations. Note that we do not have an algorithmic score for this approach as IMDb does not provide a ranking for their recommendations.

6 CONCLUSIONS & FUTURE WORK

In this paper, we analyzed and evaluated the potential of a selection of five (MF, CF, TF-IDF, Doc2Vec, NW) recommender algorithms as well as one empirical recommender approach (IMDb) to calculate narrative-driven recommendations. To be able to conduct our analyses, we crowdsourced a dataset from reddit for evaluating narrative-driven recommendations and made this dataset available



(a) *Results* (b) *Score Weights*

Figure 4: *Empirical Recommendations.* Subfigure 4a shows the results of our evaluation, comparing the empirical IMDb recommendations to the ones of the reddit community in our reddit evaluation dataset with IMDb post-filters (transparent bars) and with additional narrative-based post-filters using reddit data (colored bars). We list the different evaluation metrics on the x-axis, with the corresponding values on the y-axis, again. Subfigure 4b visualizes the best-performing score weight configuration of this experiment.

to other researchers. Moreover, we re-ranked the computed recommendation lists via post-filtering techniques based on specific user requirements from the reference dataset. With our experiments we showed that (i) all implemented recommender approaches struggle to match human-based recommendations and that (ii) the incorporation of the information contained in the narratives (e.g., in the form of post-filters) can substantially improve the performance of recommender algorithms. However, we also showed that our post-filters have to be carefully configured to maximize the benefits of the added information, as the algorithmic score is an important feature across all approaches. Particularly, when applying post-filters on empirical data, we demonstrate that our post-filtering techniques can improve existing approaches, albeit limited due to the lack of an algorithmic score.

The post-filtering techniques applied in this paper are a first step into incorporating additional information provided by users into the recommendation process. For future work, we plan to investigate other similar heuristics for comparison with the ones used in this paper and to possibly obtain a further performance improvement. Moreover, we intend to extend existing algorithms by incorporating data from reddit narratives in the training phase in the form of, for instance, additional regularization terms. This could gain insight into how fast the recommendations adjust to the given recommendation needs of a user. Currently, the recommender algorithms can not be directly trained on the reddit data due to its sparsity but, as our results show, narrative information and the previous human suggestions represent a valuable information that should be leveraged already in the training phase.

Further, we plan on applying our methods to different domains, such as books, board games, or video games, to investigate whether different communities exhibit similar or different recommendation behaviors. Moreover, we will conduct a qualitative evaluation of our recommender framework to study if our suggestions are perceived as useful by the recommendation requesters. We are also dedicated to analyze additional post-filters, informed by characteristics of our reddit evaluation dataset, as well as expanding the arsenal of implemented recommender approaches, such as deep learning and different embedding approaches for the calculation of narrative-driven recommendations. Additionally, we aim on conducting experiments on reddit, by implementing a recommender bot that users can query for recommendations, while providing a narrative. Using this bot, we will be able to evaluate the importance of additional metrics, such as diversity, serendipity or novelty in the context of narrative-driven recommendations.

In this paper we present and publish a reference evaluation dataset, as well as a first analysis of post-filtering and re-ranking strategies for incorporating narratives into recommendations. We strongly believe that our reference evaluation dataset, as well as the presented experiments in this paper will help researchers and practitioners to develop new and improve existing recommendation approaches to better tackle the problem of narrative-driven recommendations, which also represents a fundamental problem in need of novel solutions for the advance of chat and voice bots.

REFERENCES

[1] Gediminas Adomavicius, Ramesh Sankaranarayanan, Shahana Sen, and Alexander Tuzhilin. 2005. Incorporating contextual information in recommender systems using a multidimensional approach. *ACM Transactions on Information Systems (TOIS)* 23, 1 (2005), 103–145.

[2] Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE transactions on knowledge and data engineering* 17, 6 (2005), 734–749.

[3] Gediminas Adomavicius and Alexander Tuzhilin. 2011. *Context-Aware Recommender Systems.* Springer US, Boston, MA, 217–253.

[4] Gediminas Adomavicius, Alexander Tuzhilin, and Rong Zheng. 2011. REQUEST: A query language for customizing recommendations. *Information Systems Research* 22, 1 (2011), 99–117.

[5] Chumki Basu, Haym Hirsh, William Cohen, et al. 1998. Recommendation as classification: Using social and content-based information in recommendation. In *Aaai/iaai.* 714–720.

[6] Jason Michael Baumgartner. 2015. Reddit comment dataset. Website. (July 2015). https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment.

[7] Toine Bogers. 2015. Searching for Movies: An Exploratory Analysis of Movie-related Information Needs. *iConference 2015 Proceedings* (2015).

[8] Toine Bogers and Marijn Koolen. 2017. Defining and Supporting Narrative-driven Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems.* ACM, 238–242.

[9] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based N-gram Models of Natural Language. *Comput. Linguist.* 18, 4 (Dec. 1992), 467–479.

[10] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards Conversational Recommender Systems. In *KDD.* 815–824.

[11] Christina Christakou, Spyros Vrettos, and Andreas Stafylopatis. 2007. A hybrid movie recommender system based on neural networks. *International Journal on Artificial Intelligence Tools* 16, 05 (2007), 771–792.

[12] Christian Desrosiers and George Karypis. 2011. A comprehensive survey of neighborhood-based recommendation methods. *Recommender systems handbook* (2011), 107–144.

[13] Simon Funk. 2006. Netflix update: Try this at home. (2006).

[14] Sumit Ghosh, Manisha Mundhe, Karina Hernandez, and Sandip Sen. 1999. Voting for Movies: The Anatomy of a Recommender System. In *Proceedings of the Third Annual Conference on Autonomous Agents (AGENTS '99).* ACM, New York, NY, USA, 434–435.

[15] Negar Hariri, Bamshad Mobasher, and Robin Burke. 2013. Query-driven Context Aware Recommendation. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13).* ACM, New York, NY, USA, 9–16.

[16] Zan Huang, Xin Li, and Hsinchun Chen. 2005. Link Prediction Approach to Collaborative Filtering. In *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '05).* ACM, New York, NY, USA, 141–142.

[17] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 426–434.

[18] Daniel Lamprecht, Florian Geigl, Tomas Karas, Simon Walk, Denis Helic, and Markus Strohmaier. 2015. Improving Recommender System Navigability Through Diversification: A Case Study of IMDb. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business (i-KNOW '15)*. ACM, New York, NY, USA, Article 21, 8 pages.

[19] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*. 1188–1196.

[20] Tariq Mahmood and Francesco Ricci. 2009. Improving Recommender Systems with Adaptive Conversational Strategies. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia (HT '09)*. ACM, New York, NY, USA, 73–82.

[21] Harry Mak, Irena Koprinska, and Josiah Poon. 2003. Intimate: A web-based movie recommender using text categorization. In *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*. IEEE, 602–605.

[22] Lorraine McGinty and James Reilly. 2011. On the evolution of critiquing recommenders. In *Recommender Systems Handbook*. Springer, 419–453.

[23] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781* (2013).

[24] Kenta Oku, Shinsuke Nakajima, Jun Miyazaki, and Shunsuke Uemura. 2006. Context-aware SVM for context-dependent information recommendation. In *Proceedings of the 7th international Conference on Mobile Data Management*. IEEE Computer Society, 109.

[25] Umberto Panniello, Alexander Tuzhilin, Michele Gorgoglione, Cosimo Palmisano, and Anto Pedone. 2009. Experimental comparison of pre-vs. post-filtering approaches in context-aware recommender systems. In *Proceedings of the third ACM conference on Recommender systems*. ACM, 265–268.

[26] Arkadiusz Paterek. 2007. Improving regularized singular value decomposition for collaborative filtering. In *Proceedings of KDD cup and workshop*, Vol. 2007. 5–8.

[27] Patrice Perny and Jean-Daniel Zucker. 2001. Preference-based search and machine learning for collaborative filtering: the "film-conseil" movie recommender system. *Information, Interaction, Intelligence* 1, 1 (2001), 9–48.

[28] David Martin Powers. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. (2011).

[29] Paul Resnick, Neophytos Iacovou, Mitesh Suchak, Peter Bergstrom, and John Riedl. 1994. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work (CSCW '94)*. ACM, New York, NY, USA, 175–186.

[30] Ruslan Salakhutdinov and Andriy Mnih. 2007. Probabilistic Matrix Factorization. In *Nips*, Vol. 1. 2–1.

[31] Ruslan Salakhutdinov and Andriy Mnih. 2008. Bayesian probabilistic matrix factorization using Markov chain Monte Carlo. In *Proceedings of the 25th international conference on Machine learning*. ACM, 880–887.

[32] Gerard Salton and Michael J McGill. 1986. Introduction to modern information retrieval. (1986).

[33] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl. 2001. Item-based Collaborative Filtering Recommendation Algorithms. In *Proceedings of the 10th International Conference on World Wide Web (WWW '01)*. ACM, New York, NY, USA, 285–295.

[34] Paul Seitlinger, Dominik Kowald, Simone Kopeinik, Ilire Hasani-Mavriqi, Tobias Ley, and Elisabeth Lex. 2015. Attention Please! A Hybrid Resource Recommender Mimicking Attention-Interpretation Dynamics. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion)*. ACM, New York, NY, USA, 339–345.

[35] Emine Yilmaz, Evangelos Kanoulas, and Javed A. Aslam. 2008. A Simple and Efficient Sampling Method for Estimating AP and NDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '08)*. ACM, New York, NY, USA, 603–610.

3.3 Characteristics of the Microtask Workforce

This section presents two articles that analyze the characteristics of the international microtask workforce. The work presented in Section 3.3.1 maps out the socio-demographic characteristics of the international microtask workforce on the platform Figure Eight and sheds light on country-specific differences. Furthermore, it presents a cross-country comparison of the importance of microtask income in the workers' lives. Section 3.3.2 presents a new instrument for measuring the motivations of the microtask workforce, the *Multidimensional Crowdworker Motivation Scale (MCMS)*. The MCMS is the first motivation scale that was developed specifically for the context of work on microtask platforms and that offers a comprehensive representation of the motivational dimensions according to SDT. Moreover, it is the first motivation scale for the crowdworking domain that has been validated across multiple countries and income groups. Finally, Section 3.3.2 also presents the first cross-country comparison of crowdworker motivations. The studies presented in this section constitute an important step towards a better understanding of the socio-demographic characteristics and motivations of the international microtask workforce.

3.3.1 Characterizing the Global Crowd Workforce: A Cross-Country Comparison of Crowdworker Demographics

This article tackles the second research question (*RQ2*, presented in Section 1.4) by presenting a large-scale country-level study of socio-demographic characteristics of the international microtask workforce. Furthermore, the article presents an analysis of the importance of microtask income for workers in different countries.

To gain insights into the characteristics of the microtask workforces in different countries, we conducted a large survey of workers across ten countries on the platform Figure Eight. We selected the ten countries from three different World Bank income categories and additionally aimed for a high cultural diversity. To capture a diverse sample of workers in each country, we split the starting times of the surveys into three groups. In each country, 300 responses were requested during weekends, 300 were requested during typical working hours in the respective time zones and 300 were requested in the evening. We repeated the survey after eight months in order to gain insights into the stability of the distributions of the different socio-demographic characteristics. In total, we collected 18,000 responses.

Our results showed that there are substantial differences between the characteristics of the workforces in different countries, both regarding the different socio-demographic characteristics and regarding the importance of income from microtasks. Furthermore, the results showed that these characteristics were mostly stable between the two large independent samples taken at different points in time.

Characterizing the Global Crowd Workforce: A Cross-Country Comparison of Crowdworker Demographics

LISA POSCH, GESIS & GRAZ UNIVERSITY OF TECHNOLOGY

ARNIM BLEIER, GESIS

FABIAN FLÖCK, GESIS

MARKUS STROHMAIER, RWTH AACHEN UNIVERSITY & GESIS

ABSTRACT

Micro-task crowdsourcing is an international phenomenon that has emerged during the past decade. This paper sets out to explore the characteristics of the international crowd workforce and provides a cross-national comparison of the crowd workforce in ten countries. We provide an analysis and comparison of demographic characteristics and shed light on the significance of micro-task income for workers in different countries. This study is the first large-scale country-level analysis of the characteristics of workers on the platform Figure Eight (formerly CrowdFlower), one of the two platforms dominating the micro-task market. We find large differences between the characteristics of the crowd workforces of different countries, both regarding demography and regarding the importance of micro-task income for workers. Furthermore, we find that the composition of the workforce in the ten countries was largely stable across samples taken at different points in time.

1. INTRODUCTION

Doing freelance work over an online platform or by other digital means is a mode of labor that has existed since at least the 1990s. However, doing such work (i) for many, largely unknown principals, (ii) on very small-scale, non-expert tasks (often in parallel with other workers) that (iii) are continuously available from central platforms is a development that has only gained increasing popularity with an international workforce over roughly the past 15 years and has no offline equivalent. This mode of work is called micro-task crowdsourcing and, through its low-barrier nature, offers potential income opportunities for almost everyone with an internet connection.[1]

[1] Other modes of crowdsourcing exist, such as winner-takes-all contests for ideas or more macro-size tasks, essentially constituting online freelancing (cf. (Kuek et al., 2015)).

Precise estimates of the number of platforms, their users and turnovers for this type of work are hard to come by, as no official labor market statistics for crowdwork exist as of yet, and proprietary platforms seldom release such information. However, experts postulate a significant and lasting growth of microtask platforms, assuming a market size of $500 million in 2016, with the amount of global micro-task workers being put at around 9 million, up from 4 million in 2013 (Kuek et al., 2015). The World Bank (Kuek et al., 2015), the European Agency for Health and Safety at Work (European Agency for Health and Safety at Work, 2015) and other official bodies have in recent years been discussing chances and perils of this new form of income for millions of people, and they see the need for better regulation, but also plainly for better insights into the crowdsourcing market. Scholars and legislators have for instance expressed qualms about the tendency of crowdwork – often meant to offer supplementary income – to evolve into a main income source for workers in precarious economic circumstances, while at the same time being unregulated, volatile in terms of pay and availability, not offering union-typical bargaining powers and requiring predominantly monotonous work. On the flip side, opportunities through crowdwork have been highlighted, especially for inhabitants of regions with sub-par working conditions in "offline" employment (Kuek et al., 2015).

To inform this discussion of the impact of crowdwork on communities around the world, research concerned with the *demographic composition of the international crowd workforce* is very valuable, not least to enable comparisons with the more traditional, offline workforce. In this regard it is also strongly linked with the study of *why* crowdworkers are attracted to this new form of employment (e.g. Posch et al. (2017); Brewer et al. (2016); Brabham (2010)).

Further, demographic information is instrumental for optimizing the use of crowd platforms as recruiting instruments to infer knowledge about a broader ground population, or at least control for sampling biases – e.g., for using crowdworkers as an affordable and expeditious alternative in psychological testing (Paolacci and Chandler, 2014). Lastly, it is valuable for understanding task performance linked to demographic features, e.g., for labeling, translation, or speech recognition tasks (e.g. Kazai et al. (2012); Pavlick et al. (2014)).

While its useful applications seem apparent, knowledge about the demographic composition of the crowd workforce remains spotty. Out of the two mayor micro-task platforms dominating the market,[2] only the demographic composition of the predominantly American and Indian crowdworkers on Amazon Mechanical Turk (MTurk) is sufficiently well-known (e.g. Ipeirotis (2010b); Ross et al. (2010); Berg (2016)), but insights about other platforms – and particularly workers in countries outside the MTurk target audiences – are few and far between.

This paper therefore sets out to complement the existing literature by mapping out the demographics of the second micro-task market leader, CrowdFlower (since 2018 known as Figure Eight),[3] exploring its much more international crowd workforce to shed additional light on country-specific

[2] MTurk and Figure Eight (CrowdFlower) are estimated to share 80% of all revenue generated in the microtask market, with revenues approximately equal (Kuek et al., 2015).

[3] The platform's name changed from CrowdFlower to Figure Eight in 2018 and at the time of our data collection, which started in 2016, the platform's name was CrowdFlower. For consistency with the survey questions, we therefore refer to the platform as CrowdFlower rather than Figure Eight in the remainder of this paper.

differences. We conducted a survey of CrowdFlower workers in ten countries, over two time points, collecting information about their demography as well as the centrality of micro tasks in their life, regarding time spent as well as importance and use of micro-task income.

The main contributions of this paper are (1) a large-scale comparison of crowdworker demographics in ten different countries, (2) a comparison of the centrality of micro tasks in the worker's lives in these ten countries and (3) an analysis of the changes in these features between two samples taken eight months apart.

The paper is structured in the following way. Section 2 gives an overview of related work on the characteristics of crowdworkers. Section 3 describes our survey design and the process of data collection. In Section 4, we present a cross-national comparison of crowdworker demographics, and Section 5 presents a comparison of the importance that micro tasks have for workers in different countries. Finally, Section 6 concludes this paper.

2. RELATED WORK

Most research investigating demographic and economic characteristics of workers on micro-task platforms has focused on the platform **Amazon Mechanical Turk (MTurk)**. Early studies on the demographics of workers on MTurk (Ipeirotis, 2010b; Ross et al., 2009, 2010; Paolacci et al., 2010; Kazai et al., 2012) found that the vast majority of workers were located in the USA and India, and that they were young and highly educated. Workers were predominantly female in the USA and predominantly male in India. A small but significant percentage of workers relied on MTurk to make basic ends meet.

Later studies on the demographics of MTurk workers reported similar results (e.g. Goodman and Paolacci (2017); Berg (2016); Peer et al. (2017); Pavlick et al. (2014); Naderi (2018); Difallah et al. (2018)). On MTurk, American and Indian crowdworkers still constitute the vast majority of workers,[4] which is likely due to the fact that workers from other countries can only receive payment from MTurk in the form of Amazon.com gift cards (Amazon Mechanical Turk, 2016). Consistent with earlier studies, Berg (2016) found that Indian and American workers on MTurk were young and well-educated. Indian workers were predominantly male, but there was now more gender balance among workers from the US. These findings are also supported by current data collected by *mturk tracker*[5] (Ipeirotis, 2010a). Pavlick et al. (2014) conducted a study on the languages spoken by bilingual workers on MTurk and found that the majority of workers who accepted their translation tasks were located in either the USA or India. Nevertheless, there were sufficient bilingual workers to accurately and quickly complete translation tasks for 13 different languages.

Research on the demographics of workers on MTurk is closely linked with questions concerning the representativeness of MTurk samples and their suitability for different research purposes (e.g. Goodman and Paolacci (2017)). For example, Paolacci et al. (2010) compared American crowdworkers on MTurk to the general US population and found that workers in the USA were more

[4] Crowdworkers from the United States and from India constitute over 80% of the worker population on ATM (also see http://demographics.mturk-tracker.com/#/countries/all).

[5] http://www.mturk-tracker.com

representative of the population than university subject pools. Compared to the general US population, crowdworkers tended to be slightly younger and, despite being more highly educated, workers had a lower income level. This observation could be partially explained by age. Buhrmester et al. (2011) compared MTurk workers to standard Internet samples. Their MTurk sample was more diverse than both standard Internet samples and American college samples. They found that MTurk workers were similar in gender distribution, more non-white, almost equally non-American, and older than the standard Internet sample. Berinsky et al. (2012) evaluated the suitability of crowdworker samples for experimental political science and found that the respondents recruited on MTurk were more representative of the U.S. population than in-person convenience samples, but less representative than respondents recruited for Internet-based panels or national probability samples. Furthermore, they found that crowdworkers responded to experimental stimuli in a way that was consistent with prior research.

Weinberg et al. (2014) analyzed sociodemographic characteristics of workers on MTurk and compared them to the characteristics of respondents of a population-based web panel. They found that the MTurk participants were younger, more educated and there was a higher proportion of women than among the web panel participants. The MTurk sample was more divergent from the general population than the web panel. Huff and Tingley (2015) analyzed the demographics and political characteristics of MTurk workers from the United States and compared them to the respondents of the Cooperative Congressional Election Survey (CCES), a stratified sample survey conducted yearly in the United States. They found that MTurk was, in many cases, good at attracting those demographics that were difficult to attract for CCES (e.g. young Asian males). Furthermore, they found that the distribution of employment in different occupational sectors of workers on MTurk was very similar to that of CCES respondents, and that the respondents were located in similar locations on the rural-urban continuum.

Shapiro et al. (2013) investigated the suitability of crowdworker samples for conducting research on psychopathology, investigating the prevalence of different psychiatric disorders and related problems among crowdworkers on MTurk. They concluded that MTurk might be useful for studying clinical and subclinical populations. Paolacci and Chandler (2014) analyzed the characteristics of MTurk as a participant pool for the social sciences and concluded that worker samples from MTurk could replace or supplement convenience samples in psychological research, but that they should not be considered representative of a country's population.

Research on the demographics of workers on other micro-task platforms, and therefore also on workers based in countries other than the USA and India, is more scarce. Furthermore, due to reasons such as unavailability of demographic data beyond the workers' location or small sample sizes, none of these studies have so far analyzed and compared the demographics of workers at the country level.

Hirth et al. (2011) analyzed the demographics of the platform Microworkers with respect to the home countries of requesters and workers and found that the platform was much more geographically diverse than MTurk. The countries with the largest amount of workers were Indonesia, Bangladesh, India and the United States, accounting for 60% of the workforce on the platform. The remaining 40% were dispersed over a heterogeneous set of geographical locations. Using the United Nations Human Development Index to categorize countries, they found that an almost equal proportion of workers were located in countries with low development (24%) and countries with very high development (21%), while the majority of the workforce was located in countries with medium development (45%).

Martin et al. (2017) compared the demographics of workers on MTurk to the demographics of workers on the platforms Microworkers and Crowdee. The study grouped workers on the Microworkers platform into workers located in “Western countries” (including all workers from Europe, Oceania and North America) and workers located in “developing countries” (including all workers from South America, Asia and Africa). The results of their demographic survey indicated that workers on Microworkers and Crowdee were predominantly male, younger than workers on MTurk and highly educated. A large proportion of workers reported working either full-time or part time on all three platforms. Regarding the differences between “Western countries” and “developing countries,” the study found that workers in the “developing countries” group were younger, lived in larger households, were more educated, had a lower household income and spent more time on the platform, compared to workers in the “Western countries” group.

Further, few studies have concerned themselves with the workforce demographics of the platform CrowdFlower, which covers half the market share for micro-tasks and traditionally employs a geographically diverse set of workers.

Berg (2016) collected demographic data from a geographically diverse sample of workers on CrowdFlower and found that only 2.8% of workers (10 respondents) were from the US and 8.5% were from India (30 respondents). The workers on CrowdFlower were predominantly male and more educated than American workers on MTurk, but less educated than Indian workers on MTurk. Peer et al. (2017) examined the demographics of workers on CrowdFlower and Prolific Academic and compared them to the demographics of workers on MTurk. The study found that, compared to MTurk, both CrowdFlower and Prolific Academic had a higher proportion of male workers and the mean age was similar on all three platforms. CrowdFlower had the highest diversity in terms of race and both CrowdFlower and Prolific Academic were much more diverse in worker location than MTurk. On all three platforms, workers were highly educated.

In sum, there have been extensive studies on the characteristics and demographic composition of American and Indian workers on MTurk, whereas research on the characteristics of crowdworkers on other micro-task platforms and on the demographic composition and characteristics of workers in countries other than the USA and India remains sparse. In comparison, we provide a *survey of workers pre-selected to cover similar respondent numbers for ten diverse countries, over two time points*, whereas previous studies have studied samples that were not stratified by locations, and have mostly not controlled for temporal changes. In doing so, *we have conducted the most comprehensive scientific collection of worker characteristics on CrowdFlower to date, with 11,946 individual responses collected* (after spam removal). Using the data collected with our survey, we provide the first country-level comparison of (i) the demographics and (ii) the centrality of micro tasks in the lives of crowdworkers on this platform and show that notable differences exist between countries.

3. SURVEY

In order to provide insights into the characteristics of the international crowd workforce, we conducted a large survey in ten different countries and at two points in time on the CrowdFlower platform.

3.1. Data Collection

We posted the survey as a micro task on CrowdFlower. The task included seven questions about the workers' demographics and three questions about the centrality of micro tasks in the workers' lives. Furthermore, the task contained questions about the workers' motivation for putting effort into micro tasks, which were used for the validation of the Multidimensional Crowdworker Motivation Scale (Posch et al., 2017). Anonymity was ensured in the task instructions.

We collected data of workers from ten countries, with 900 participants in each country at each time point. In our country selection, we aimed for diverse income levels by selecting countries from three different World Bank income groups.[6] Furthermore, we aimed for a high cultural diversity and

[6] The World Bank country classification is available at http://databank.worldbank.org/data/download/site-content/CLASS.xls. Here, we use the group label "middle Income" (MID) for the upper middle income group

***Table 1.** Sample sizes and percentage of spam received. This table shows the sample sizes of the different groups at both time points, as well as the percentage of spam received.* $\mathbf{N}_{raw}$ *shows the total number of responses collected, before removing spam.* $\mathbf{N}_{T1}$ *and* $\mathbf{N}_{T2}$ *show the number of responses after spam removal in sample T1 and T2, respectively.* $\mathbf{Spam}_{T1}$ *and* $\mathbf{Spam}_{T2}$ *show the percentage of workers who did not pass all attention checks, for T1 and T2.*

Group	Code	$\mathbf{N}_{raw}$	$\mathbf{Spam}_{T1}$	$\mathbf{N}_{T1}$	$\mathbf{Spam}_{T2}$	$\mathbf{N}_{T2}$
All	ALL	18000	35 %	5857	32%	6089
High Income	HIGH	5400	28 %	1952	26%	1988
Middle Income	MID	5400	32 %	1834	31%	1863
Low Income	LOW	5400	44 %	1508	38%	1679
USA	USA	1800	20 %	721	14%	776
Spain	ESP	1800	25 %	677	30%	634
Germany	DEU	1800	38 %	554	36%	578
Brazil	BRA	1800	45 %	496	43%	509
Russia	RUS	1800	25 %	677	21%	708
Mexico	MEX	1800	27 %	661	28%	646
India	IND	1800	32 %	608	28%	645
Indonesia	IDN	1800	55 %	401	47%	476
Philippines	PHL	1800	45 %	499	38%	558
Venezuela	VEN	1800	37 %	563	38%	559

sufficient activity on CrowdFlower.[7] The countries that we selected for the high income group were USA, Germany and Spain, for the middle income group we selected Brazil, Russia and Mexico, and for the low income group we selected India, Indonesia and the Philippines. Additionally, we collected data from Venezuela because it was the most active country on CrowdFlower at the time of the start of the data collection. However, Venezuela represents a special case concerning income due to the circumstance that the black market exchange rate deviates from the official exchange rate to a large extent (Bloomberg News, 2016). Therefore, we did not include Venezuela in any of our income groups.

We posted the survey on CrowdFlower at different times during the day and the week, in order to capture a diverse sample of workers.[8] For each country, 300 responses were requested during typical working hours (8:00 am to 5:00 pm in the appropriate time zone), 300 responses were requested in the evening (6:00 pm to 11:00 pm in the appropriate time zone), and 300 responses were requested during weekends. After the first round of data collection, which took place in October and November 2016 (`T1`), we conducted a second round of data collection in June and July 2017 (`T2`). The survey was conducted in English in all countries. While this approach only captures crowdworkers with sufficient English skills, demand for crowdworkers is driven by Anglophone clients and English is the dominant language in task requests (Kuek et al., 2015). Congruently, English is expected by CrowdFlower to be spoken by all workers at a sufficient level to solve tasks, as made apparent by its interface language and English being assumed a guaranteed language skill for all workers in the platform's worker language selection settings.

and "low Income" (LOW) for the lower middle income group for better readability.

[7] The country had to either be high in the Alexa (http://www.alexa.com/) ranking or one of the top contributing countries in at least one of CrowdFlower's partner channels.

[8] There are indications that worker composition varies by time of the day and day of the week, see e.g. http://demographics.mturk-tracker.com

***Table 2.* Survey Questions. This table shows the survey questions in our CrowdFlower task.**

Demographics	
D1	*What is your gender?*
D2	*What is your age?*
D3	*What is your marital status?*
D4	*How many people live in your household?*
D5	*What is your highest education level?*
D6	*What is your employment status (CrowdFlower tasks excluded)?*
D7	*What is your approximate household income, per year (after taxes, in US$)?*
Importance of Micro Tasks	
I1	*How much time do you spend on CrowdFlower, per week?*
I2	*Is the money from CrowdFlower your primary source of income?*
I3	*What do you do with the money that you earn on CrowdFlower?*

In the tasks, we included four attention checks for detecting spam, such as workers clicking randomly or accepting the task despite having insufficient English skills.[9] Table 1 shows the number of respondents per income group and country for each time point, as well as the percentage of spam

***Table 3.* Exemplary responses to the open ended survey question "What do you do with the money that you earn on CrowdFlower?".**

Basic Expenses "I buy food!!" (USA), "I use the money to help pay my monthly rent." (USA), "buy sensors to glucose measure" (Spain), "I use it for my daily needs like to pay rent and buy my essentials." (India), "use it for my medicines" (India)
Leisure Activities "I will use to pay for my hobbies." (USA), "entertainment, eating out" (USA), "Go to the cinema." (Spain), "With the money I usually do trips." (Spain), "Use it as pocket money" (India)
Save/Invest "Put it in a savings account" (USA), "Build a BitCoin investment portfolio" (USA), "I keep it for the future" (Spain)", "I save all the money I earn" (Spain), "i save the money for investments" (India), "saving for marriage and future life" (India)
Buy Gifts "I will save it to try and afford a gift for my children for christmas" (USA), "spend it on xmas presents for my kid" (USA), "small Gifts" (Spain), "buy gifts to my four daughters" (Spain), "Use it to buy gifts for my children." (India), "used it for my mom dad's anniversary" (India)
Education "Pay for my college tuition." (USA), "Use it for my driving test." (Spain), "Save it [t]o pay for my college expenses" (Spain), "The Money I have Earned in CrowdFlower is Used for my Studies." (India), "for further studies" (India)
Donate to Charity "I will spend for my family and remaining to charity." (India), "I want to do lot of the things, primary is to donate a share out of it [...]." (India), "Almost 80% paid for poor children's fee." (India), "helping to poor peoples" (India)
Other "Nothing yet, this is my first task." (USA), "Multiple things, nothing in particular." (Spain), "it is very small amount to spend i have not earned so much" (India), "I have not much enough to withdraw it." (India)

[9] For a detailed description of the spam filtering process, see (Posch et al., 2017).

received and the number of respondents after spam removal. As it is the crowdworkers' choice whether to accept a task or not, our samples are necessarily self-selected, as is generally the case for surveys on micro-task platforms.

In the survey, we asked crowdworkers about seven demographic characteristics and about three aspects concerning the importance of micro tasks in their life. Table 2 shows the questions. The question about crowdworkers' use of the money earned through micro tasks (I3) was constructed as a multiple choice question. We aimed for a high-level distinction of money use to keep the number of answer options low (and reduce the total survey length). As standard survey instruments for capturing expenditures of households or individuals are very detailed in their classifications and do not provide canonical distinctions at a sufficiently high level for our purposes[10] – and to account for potential particularities of crowdworker money use patterns – we opted for an inductive approach to constructing the answer options.

To this end, we posted a preceding open ended survey task, where we asked workers the question *"What do you do with the money that you earn on CrowdFlower?"* For answering the question, we provided workers with a free text field for their answer, which could be arbitrarily long. We posed this question to workers in the USA, Spain and India.[11] In each country, 300 workers were surveyed in October 2016. Two authors of this paper then manually categorized the open-ended responses. Workers often reported more than one use for the money earned through micro tasks, so each answer could be coded with multiple categories.

We then used these manually identified categories to construct the answer options for the survey question I3: (1) *I use the money for basic living expenses (food, rent, sanitary items, medical care,...)*, (2) *I spend the money on leisure activities (hobbies, games, holidays, sports,...)*, (3) *I save/invest the money*, (4) *I use the money to buy gifts for other people*, (5) *I use the money to finance my education*, (6) *I donate the money to charity*, and (7) *Other purposes*. Table 3 shows example answers for each category, along with the country the answers stem from.

4. DEMOGRAPHICS

In this section, we report the results of the demographics section of the survey (see questions D1-D7 in Table 2). For each demographic characteristic, we report the proportion of each answer choice in the ten countries as an average of T1 and T2, the differences in proportion between the countries and, per country, the differences in proportions between the two samples taken eight months apart. As a measure of difference, we report the Jensen-Shannon (JS) divergence (Lin, 1991) between the respective answer distributions for each demographic characteristic. The JS divergence quantifies how dissimilar two distributions are and is bounded by 1 and 0. A value of 0 indicates equivalence

[10] Cf., e.g., the "Consumer Expenditure Survey Interview Questionnaire" (U.S. Bureau of Labor Statistics, 2018) or the "Classification of Individual Consumption according to Purpose" (United Nations Department Of Economic And Social Affairs - Statistics Division, 2000). If coarser distinctions are defined, they are generally at the binary consumption vs. savings/other expenditures level (cf. (Destatis, 2013)).

[11] For the development of the answer options, we used the same countries as for the development of the Multidimensional Crowdworker Motivation Scale (Posch et al., 2017). USA and India were selected because these countries have significant populations of crowdworkers on different platforms. Spain was selected in order to include a European country with a sufficiently large population of crowdworkers.

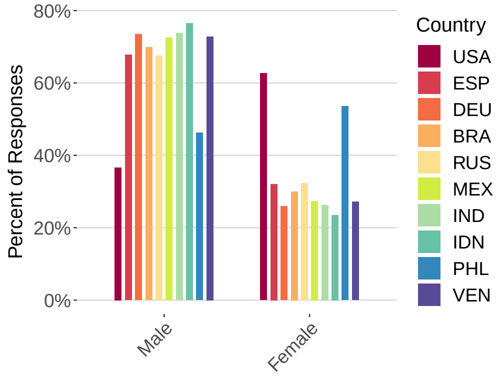


***Figure 1.* Gender Distribution. This figure shows the gender distribution of workers in the different countries. The bar height represents the average of `T1` and `T2`.**

between the distributions and higher values indicate the degree of dissimilarity. The reported JS divergences between two countries are the averages of the divergences between these countries at T1 and T2.

4.1. Gender

In most countries, crowdworkers were predominantly male, with the proportion of male workers exceeding 60%. The gender distribution was similar in all countries with the exceptions of the USA and the Philippines, which were the only two countries where female workers constituted the majority. The most gender balanced workforce was present in the Philippines, with 52% (in T1) and 55% (in T2) percent of workers being women. Figure 1 shows the gender distribution in the ten countries. The height of the bars corresponds to the average of the proportions at T1 and T2.

The answer options to the gender question included a third category, "other," which is not included in Figure 1 due to the small number of responses. The differences between the sums of the male and female percentages and 100% are due to this third category.

The gender distributions of American and Indian workers are consistent with findings of early studies on MTurk (e.g. Ipeirotis (2010b); Paolacci et al. (2010)) which found that in the United States, there were more female than male workers, and in India, there were more male workers. However, the United States crowd workforce on MTurk, at the time of data collection, was more gender balanced than the US-based crowd workforce on CrowdFlower.[12] Ipeirotis (2010b) hypothesized that this gender distribution difference between India and the United States may be due to the fact that in the United States, MTurk is often used by stay-at-home parents and underemployed or unemployed workers (which are more likely to be female), while in India, workers are more likely to rely on

[12] For data on the gender distribution of American and Indian workers on MTurk, see http://demographics.mturk-tracker.com/#/gender/all.

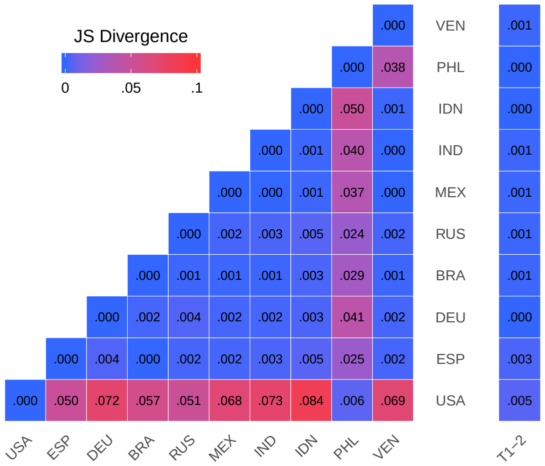


***Figure 2.* Gender JS. This figure shows the JS divergences between the gender distributions of the different countries. The bar on the right shows the JS divergence between T1 and T2 for each individual country.**

MTurk as a primary source of income. However, our results show that this does not generally hold true for the differences between the gender distribution of high income and low income countries.

Figure 2 shows the JS divergences of the answer distributions between each country pair and, for each country, the divergence between T1 and T2. The gender distribution was mostly stable between the time points, with Spain and the USA exhibiting the largest differences in distributions. The divergence in the USA was mainly due to the gender category “other,” as which ten crowdworkers identified in T2, compared to none in T1. The divergence in Spain was due to an increased proportion of female workers in T2, where 35% of workers reported being female in T2 compared to 30% in T1. While the change was less pronounced in other countries, the percentage of female crowdworkers slightly increased from T1 to T2 in all countries except Russia.

4.2. Age

Crowdworkers were young in all countries, with most crowdworkers being between 18 and 34 years of age. This is consistent with studies on MTurk (e.g. Ipeirotis (2010b); Ross et al. (2010); Berg (2016)), which found that younger workers were overrepresented on the platform.

The country with the oldest population of crowdworkers on CrowdFlower was Russia, which had by far the lowest proportion of workers aged between 18-24 years and the highest population of workers aged between 35 and 54 years old. Venezuela had the highest proportion of very young workers (aged 18-24). Figure 3 shows the age distribution in the ten countries. Data from mturk

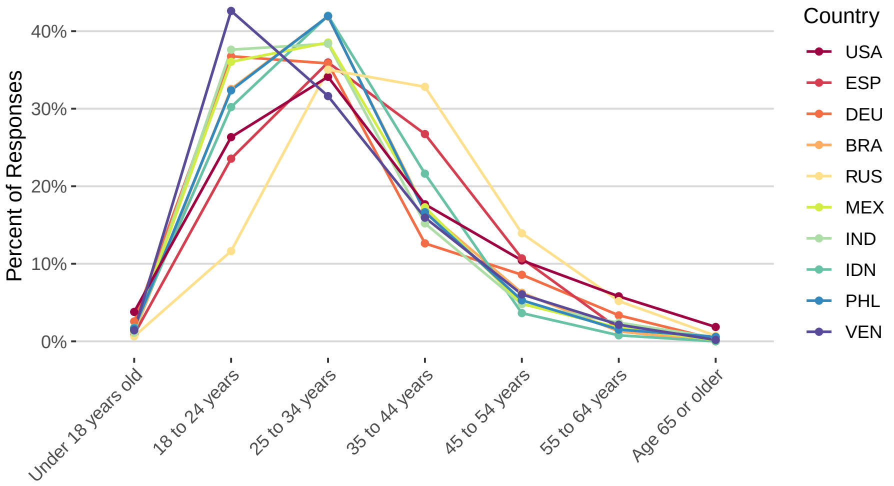
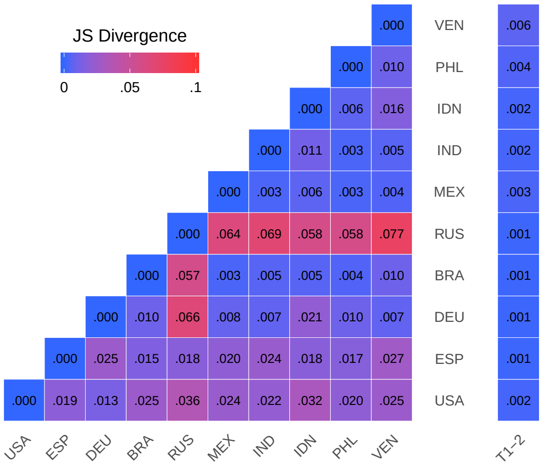


Figure 3. ***Age Distribution. This figure shows the age distribution of workers in the different countries. The percentages represent the average of*** `T1` ***and*** `T2`***.***



	USA	ESP	DEU	BRA	RUS	MEX	IND	IDN	PHL	VEN	T1-2
VEN										.000	.006
PHL									.000	.010	.004
IDN								.000	.006	.016	.002
IND							.000	.011	.003	.005	.002
MEX						.000	.003	.006	.003	.004	.003
RUS					.000	.064	.069	.058	.058	.077	.001
BRA				.000	.057	.003	.005	.005	.004	.010	.001
DEU			.000	.010	.066	.008	.007	.021	.010	.007	.001
ESP		.000	.025	.015	.018	.020	.024	.018	.017	.027	.001
USA	.000	.019	.013	.025	.036	.024	.022	.032	.020	.025	.002

Figure 4. ***Age JS. This figure shows the JS divergences between the age distributions of the different countries. The bar on the right shows the JS divergence between*** `T1` ***and*** `T2` ***for each individual country.***

tracker[13] indicates that Indian workers on MTurk tend to be younger than American workers. This difference also seems to be present on CrowdFlower, especially for the proportion of workers aged 18 to 24 years, which formed a much higher percentage of the Indian crowd workforce than the American crowd workforce.

Figure 4 shows the JS divergences of the age distributions. In most countries, there was little difference in age distribution between `T1` and `T2`. Venezuela had the largest difference in age distribution between the two time points, mostly due to an increase in young workers in the age bracket 18 to 24 years and a decrease in workers aged over 24.

4.3. Marital Status

Most countries had a higher proportion of non-married workers than married workers. Of all countries, Russia had the highest proportion of married crowdworkers and it was the only country with more than 60% married workers. USA and Russia were the countries with the largest proportion of divorced or separated workers. The countries with the highest proportion of non-married workers were Germany, the Philippines and Venezuela. Figure 5 shows distribution of the workers' marital status in the ten countries. The response option for this survey question also included the category "widowed," which received a very small number of responses and is therefore not included in Figure 5.

Figure 6 shows the JS divergences of the answer distributions. Russia had the highest divergences with other countries due to the high proportion of married workers. Regarding the differences between the time points, in most countries there was a slight decrease in the proportion of married workers from `T1` to `T2`. The only country where this was not the case was Germany, which also had the most stable distribution of marital status between the time points.

4.4. Household Size

Germany had the highest proportion of single and two-person households, followed by the USA. All other countries had a very low proportion of single households (below 10%). The Philippines was the country with by far the highest proportion of households with more than seven persons, with more than double the proportion of all other countries. Spain had the highest proportion of four-people households, and Russia had the highest proportion of three-person households. India's crowd workforce reported the lowest proportion of single households. Figure 7 shows the distribution of household size.

Data from mturk tracker[14] shows that on MTurk, workers in India tend to live in larger households than American workers. This difference in household size was also present in the workers on CrowdFlower. Workers located in the United States mainly lived in households with two or three persons, while a four-person household was the most common response among workers in India.

Figure 8 shows the JS divergences of the household size distributions. Generally, we found the largest divergences between the countries of the high income group as well as Russia and the low income countries. There were no large differences in household size distribution between the two

[13] http://demographics.mturk-tracker.com/#/yearOfBirth/all

[14] http://demographics.mturk-tracker.com/#/householdSize/all

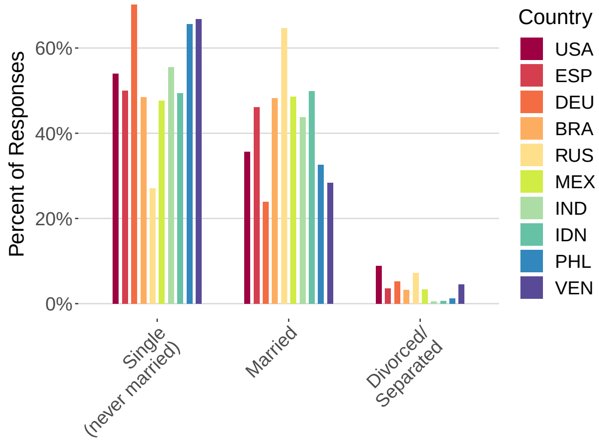


Figure 5. ***Distribution of Marital Status. This figure shows the marital status distribution of workers in the different countries. The bar height represents the average of T1 and T2.***

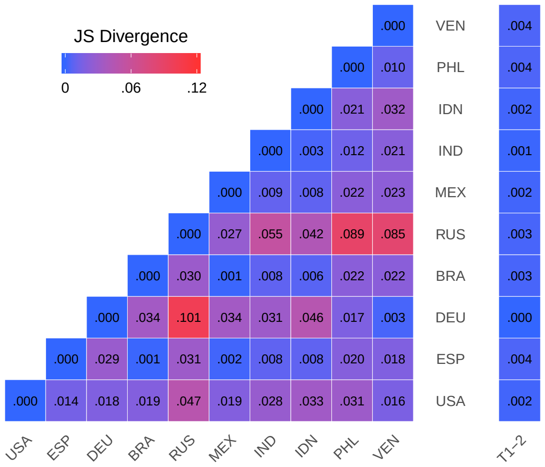


Figure 6. ***Marital Status JS. This figure shows the JS divergences between the marital status distributions of the different countries. The bar on the right shows the JS divergence between T1 and T2 for each individual country.***

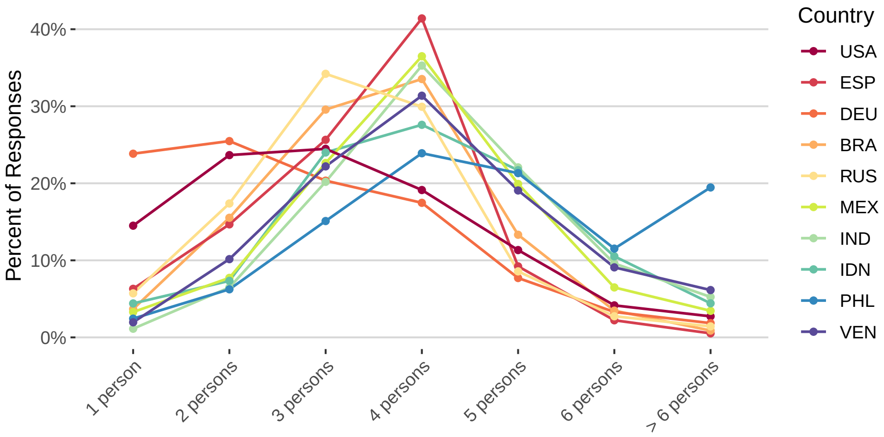
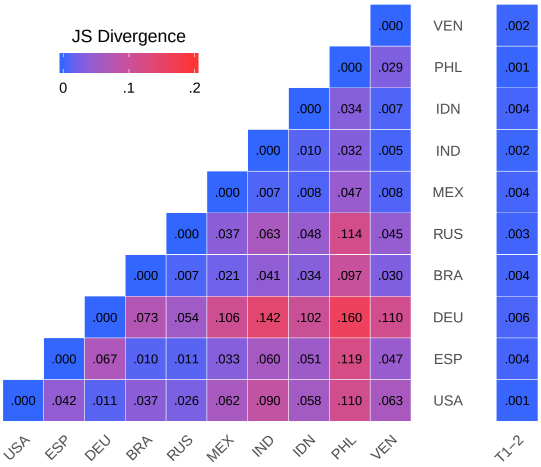


Figure 7. Distribution of Household Size. This figure shows the household size of workers in the different countries. The percentages represent the average of T1 and T2.



	USA	ESP	DEU	BRA	RUS	MEX	IND	IDN	PHL	VEN	T1–2
VEN										.000	.002
PHL									.000	.029	.001
IDN								.000	.034	.007	.004
IND							.000	.010	.032	.005	.002
MEX						.000	.007	.008	.047	.008	.004
RUS					.000	.037	.063	.048	.114	.045	.003
BRA				.000	.007	.021	.041	.034	.097	.030	.004
DEU			.000	.073	.054	.106	.142	.102	.160	.110	.006
ESP		.000	.067	.010	.011	.033	.060	.051	.119	.047	.004
USA	.000	.042	.011	.037	.026	.062	.090	.058	.110	.063	.001

Figure 8. Household Size JS. This figure shows the JS divergences between the household size distributions of the different countries. The bar on the right shows the JS divergence between T1 and T2 for each individual country.

time points. The German sample had the largest difference, with more workers reporting living in two-person households in T2 than in T1, and less workers reporting three-person households.

4.5. Employment Status

The question regarding workers' employment status asked crowdworkers to explicitly exclude their activity on CrowdFlower. Figure 9 shows the distribution of employment status.

In almost all countries, over 35% workers had a full-time job besides their activity on CrowdFlower. The only exception to this was Venezuela, where only 28% had full-time jobs at T1 besides CrowdFlower. This percentage was even lower in T2, where only 23% of Venezuelan workers reported having full-time jobs. A significant proportion of workers reported being in education, with Germany and Venezuela having the highest proportion of workers in education. The highest proportion of unemployed workers was reported in the United States, followed by Venezuela. Very few workers reported being retired, which is very likely due to the overall young age of the workers.

Figure 10 shows the JS divergences between the employment status distributions. The largest difference in employment status distribution was between Russia and Venezuela. While Russian workers reported the highest percentage of workers in full-time employment, Venezuela reported the lowest percentage of all countries. Furthermore, there were large differences between Russia and Venezuela in the proportion of workers who reported being unemployed or in education.

In most countries, less workers reported working full-time in T2 than in T1, while the percentage of unemployed workers, workers in education and part-time workers increased from T1 to T2. In Brazil, which had the largest JS divergence between the time points, this change was most pronounced, with a large decrease of workers in full-time employment (from 59.5% in T1 to 46.2% in T2) and a large increase of unemployed workers (from 13.1% in T1 to 21.6% in T2). An exception to this pattern was Germany, where the percentage of workers employed full-time stayed roughly the same, while there was a slight decrease in unemployed workers and a slight increase of workers in education. The second-largest JS divergence between time points was in Indonesia, where a lower proportion of workers reported having a full-time job in T2 than in T1, and a higher proportion of workers reported holding a part-time job.

4.6. Education Level

Crowdworkers on CrowdFlower are generally well educated. The proportion of workers having a Bachelor's degree or higher was 30% or above in all countries. Figure 11 shows the distribution of education level.

Workers in the low income group countries reported especially high education levels. The countries with the highest proportion of college graduates were India and the Philippines. India also had the highest proportion of workers with a Master's degree of all countries.

Our finding that crowdworkers are generally highly educated is consistent with the findings of studies on the demographics of MTurk (e.g. Berg (2016)), and it contrasts with the notion that microwork is especially attractive to unemployed people with no specialized skills (e.g. Kuek et al. (2015)). The fact that workers from lower income countries tend to have higher education levels is consistent with the findings of studies on MTurk (e.g. Ipeirotis (2010b)), which found that Indian

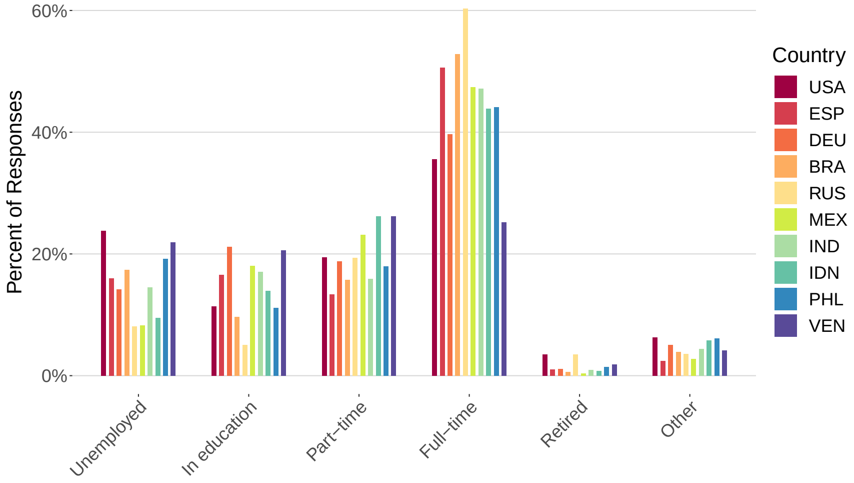
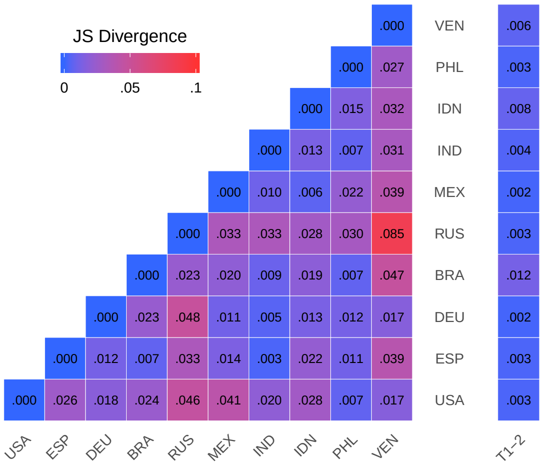


Figure 9. Employment Status Distribution. This figure shows the employment status distribution of workers in the different countries, CrowdFlower tasks excluded. The bar height represents the average of T1 and T2.



	USA	ESP	DEU	BRA	RUS	MEX	IND	IDN	PHL	VEN	T1–2
VEN										.000	.006
PHL									.000	.027	.003
IDN								.000	.015	.032	.008
IND							.000	.013	.007	.031	.004
MEX						.000	.010	.006	.022	.039	.002
RUS					.000	.033	.033	.028	.030	.085	.003
BRA				.000	.023	.020	.009	.019	.007	.047	.012
DEU			.000	.023	.048	.011	.005	.013	.012	.017	.002
ESP		.000	.012	.007	.033	.014	.003	.022	.011	.039	.003
USA	.000	.026	.018	.024	.046	.041	.020	.028	.007	.017	.003

Figure 10. Employment Status JS. This figure shows the JS divergences between the employment status distributions of the different countries. The bar on the right shows the JS divergence between T1 and T2 for each individual country.

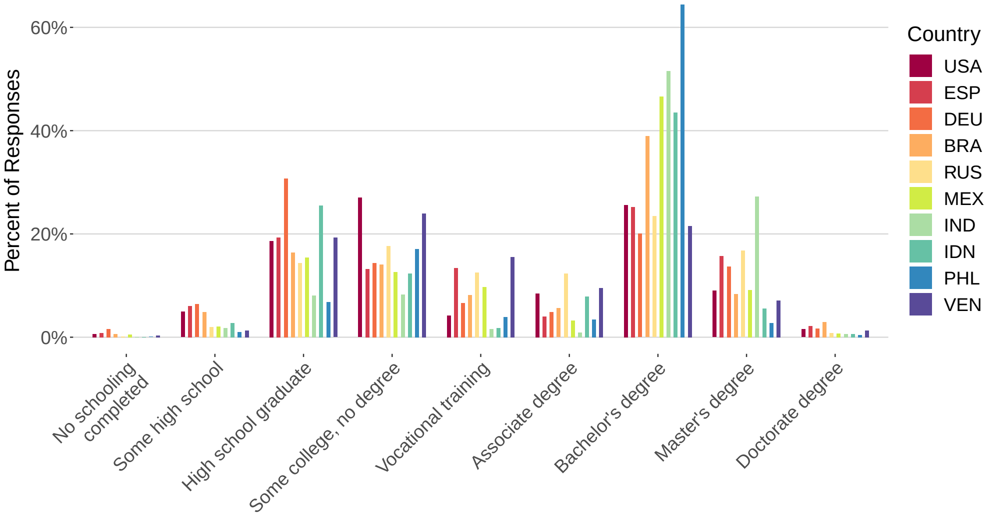
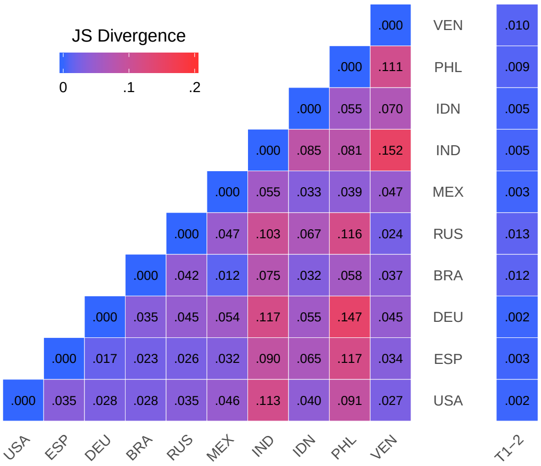


Figure 11. ***Education Level Distribution. This figure shows the education level distribution of workers in the different countries. The bar height represents the average of*** `T1` ***and*** `T2`***.***



	USA	ESP	DEU	BRA	RUS	MEX	IND	IDN	PHL	VEN	T1–2
VEN										.000	.010
PHL									.000	.111	.009
IDN								.000	.055	.070	.005
IND							.000	.085	.081	.152	.005
MEX						.000	.055	.033	.039	.047	.003
RUS					.000	.047	.103	.067	.116	.024	.013
BRA				.000	.042	.012	.075	.032	.058	.037	.012
DEU			.000	.035	.045	.054	.117	.055	.147	.045	.002
ESP		.000	.017	.023	.026	.032	.090	.065	.117	.034	.003
USA	.000	.035	.028	.028	.035	.046	.113	.040	.091	.027	.002

Figure 12. ***Education Level JS. This figure shows the JS divergences between the education level distributions of the different countries. The bar on the right shows the JS divergence between*** `T1` ***and*** `T2` ***for each individual country.***

workers on MTurk tend to have more education than workers from the United States. An exception to this pattern seems to be Venezuela, where workers tend to be less educated than in other low income countries.[15]

Very few workers reported having no schooling completed at all (below 2% in all countries) and only a small proportion of workers reported having only "some high school." Germany had the highest proportion of workers with a high school degree but no college education.

Figure 12 shows the JS divergences between the education level distributions. The largest difference in distribution was between Venezuela and India, with the proportion of Indian workers with a Bachelor's degree being more than twice as high and the proportion of workers with a Master's degree being over three times higher than the proportion in Venezuela.

Regarding the difference between the two time points, we found the largest differences in Russia, Brazil and Venezuela. Russia had less workers with a Bachelor's or Master's degree in T2 than in T1. In Brazil, the proportion of workers reporting a high school degree but no college degree increased and the proportion of workers reporting some high school, a Bachelor's degree or associate degree decreased. In Venezuela, the proportion of high school graduates with no college increased, while the proportion of workers reporting vocational training or an associate degree decreased.

4.7. Yearly Household Income

In order to meaningfully capture household income in a set of countries with wildly varying average incomes, we created logarithmic bins[16] for the response options. The question asked workers to report an estimate of their annual disposable household income (i.e. after taxes) in US dollars. Figure 13 shows the household income distribution for each country.

Workers from Venezuela, while classified as an "upper middle income" country by the World Bank, reported by far the lowest annual household income. This supports our decision to not include Venezuela in any of the income groups. Apart from Venezuela, the reported income distributions are largely consistent with the World Bank classification of the countries, with the United States, Spain and Germany reporting higher incomes (despite the smaller reported household size) and India, Indonesia and the Philippines reporting lower incomes. Unsurprisingly, data from mturk tracker[17] shows that on MTurk, Indian workers also tend to report lower household incomes then workers from the United States.

While the proportion of workers reporting an annual income below US$ 3,000 was much higher in low income countries than in high income countries, a significant proportion of workers in high income countries also reported a yearly house income of less than US$ 3,000. There might be several explanations for this, such as students living on student loans, unemployed workers living off their savings, or workers on welfare benefits who do not consider the benefits as "income."

Figure 14 shows the differences between the household income distributions. The largest differences

[15] While we did not include Venezuela in the low income country group due to the reasons stated in Section 3, Venezuelan workers reported a very low household income.

[16] We rounded the logarithmically spaced numbers for better readability in the answer options.

[17] http://demographics.mturk-tracker.com/#/householdIncome/all

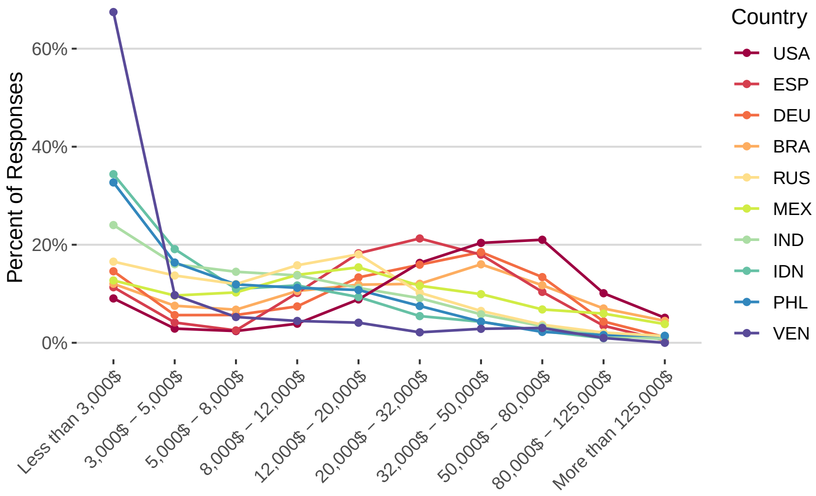
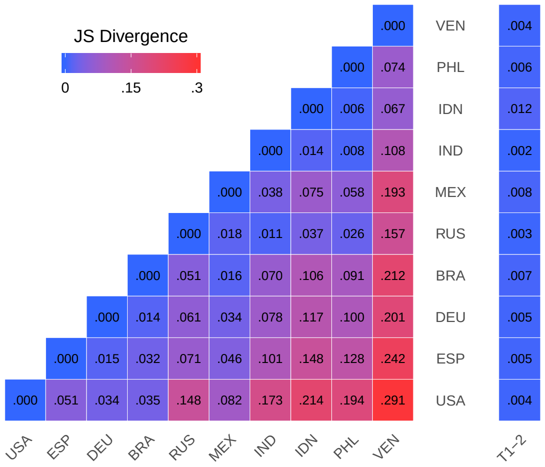


Figure 13. ***Distribution of Household Income. This figure shows the household income distribution of workers in the different countries. The percentages represent the average of*** `T1` ***and*** `T2`***.***



	USA	ESP	DEU	BRA	RUS	MEX	IND	IDN	PHL	VEN	T1–2
VEN										.000	.004
PHL									.000	.074	.006
IDN								.000	.006	.067	.012
IND							.000	.014	.008	.108	.002
MEX						.000	.038	.075	.058	.193	.008
RUS					.000	.018	.011	.037	.026	.157	.003
BRA				.000	.051	.016	.070	.106	.091	.212	.007
DEU			.000	.014	.061	.034	.078	.117	.100	.201	.005
ESP		.000	.015	.032	.071	.046	.101	.148	.128	.242	.005
USA	.000	.051	.034	.035	.148	.082	.173	.214	.194	.291	.004

Figure 14. ***Yearly Household Income JS. This figure shows the JS divergences between the household income distributions of the different countries. The bar on the right shows the JS divergence between*** `T1` ***and*** `T2` ***for each individual country.***

in household income were generally found between the countries in the low income group (and Venezuela) and the countries in the high income group, with the largest difference being between the USA and Venezuela.

Between `T1` and `T2`, the household income distributions remained largely stable. We observed the largest change in Indonesia, where in `T2` more workers reported a yearly income below US$ 3,000 (39%) than in `T1` (30%). The second largest change between the time points was in Mexico, where the number of workers reporting a household income between US$32,000 and US$50,000 decreased from 12% to 7% while the proportion of workers reporting a lower income increased.

5. IMPORTANCE OF MICRO TASKS FOR CROWDWORKERS

In this section, we compare the importance of micro tasks and micro-task income for workers in the ten different countries. Our survey included three questions about different aspects concerning the centrality of micro tasks in the workers' lives (see questions I1-I3 in Table 2). Analogously to the previous section, we report the proportion of each answer choice in the ten countries as an average between `T1` and `T2` as well as the JS divergences (Lin, 1991) of the answer distributions between the countries and between the two time points.

5.1. Weekly Time Spent on CrowdFlower

Figure 15 shows, for the ten countries, how much time workers report spending on CrowdFlower per week. Venezuela, the Philippines and Indonesia were the countries with the highest proportion of workers who reported spending more than 20 hours per week on CrowdFlower and Venezuela had the highest proportion of workers spending more then 40 hours per week on the platform. In all countries, but especially in the countries in the high and middle income groups, there was a significant proportion of workers who used CrowdFlower less than two hours per week. The countries in the high income group had the highest proportion of workers who reported spending less than one hour per week on CrowdFlower.

Figure 16 shows the JS divergences between the answer distributions. Regarding the differences between countries, countries in the high income group were generally most dissimilar to countries in the low income group, with countries in the low income group generally spending more time on the platform.

The largest change in distribution between `T1` and `T2` was in Venezuela. In `T2`, the proportion of Venezuelan workers spending over 40 hours per week on CrowdFlower (19.5%) was almost double the proportion reported in `T1` (9.6%). This increase was likely due to changes in the economic situation of the country, making CrowdFlower an increasingly attractive source of income for Venezuelan workers.

5.2. Dependency on Micro-Task Income

Crowdworkers in countries of the high and middle income groups reported the lowest percentages of reliance on CrowdFlower as their primary source of income. There were no large differences between the countries of the high income group and those of the middle income group, and the lowest reliance on CrowdFlower as a main source of income was in Russia, a country in the middle

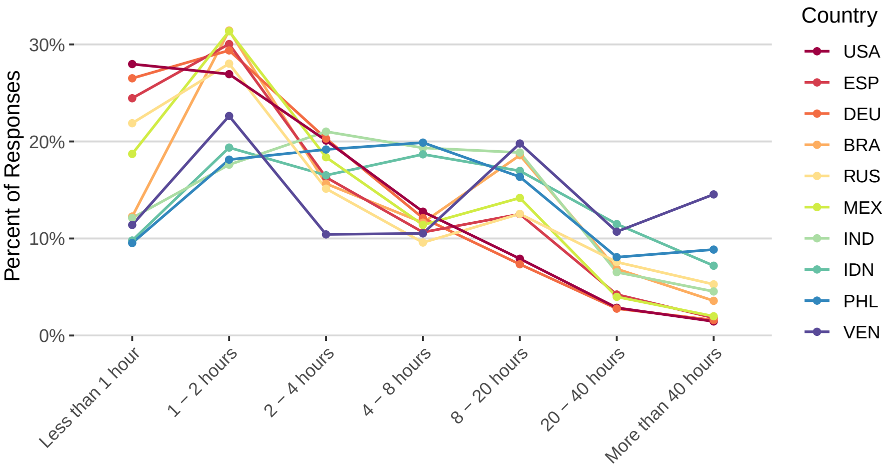
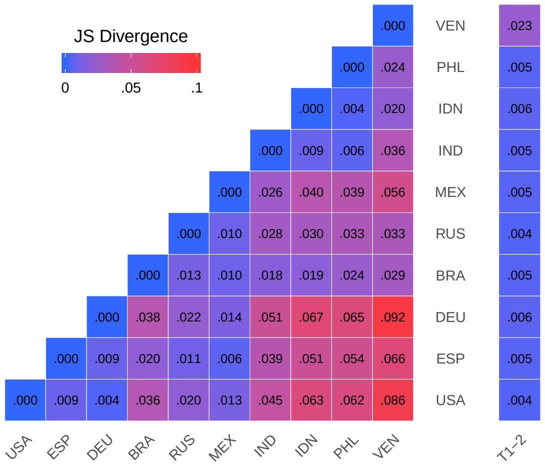


Figure 15. ***Time Spent on CrowdFlower per Week. This figure shows the distribution of weekly time spent on the platform by workers in the different countries. The percentages represent the average of*** `T1` ***and*** `T2`***.***



	USA	ESP	DEU	BRA	RUS	MEX	IND	IDN	PHL	VEN	T1–2
VEN										.000	.023
PHL									.000	.024	.005
IDN								.000	.004	.020	.006
IND							.000	.009	.006	.036	.005
MEX						.000	.026	.040	.039	.056	.005
RUS					.000	.010	.028	.030	.033	.033	.004
BRA				.000	.013	.010	.018	.019	.024	.029	.005
DEU			.000	.038	.022	.014	.051	.067	.065	.092	.006
ESP		.000	.009	.020	.011	.006	.039	.051	.054	.066	.005
USA	.000	.009	.004	.036	.020	.013	.045	.063	.062	.086	.004

Figure 16. ***JS of Weekly Time Spent on CrowdFlower. This figure shows the JS divergences between the answer distributions of the different countries. The bar on the right shows the JS divergence between*** `T1` ***and*** `T2` ***for each individual country.***

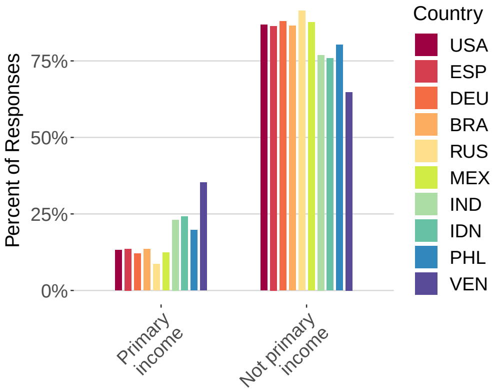


Figure 17. Dependency on CrowdFlower Income. This figure shows the proportion of workers who reported micro-task income being their primary/non-primary source of income in the different countries. The bar heights represent the averages of `T1` *and* `T2`*.*

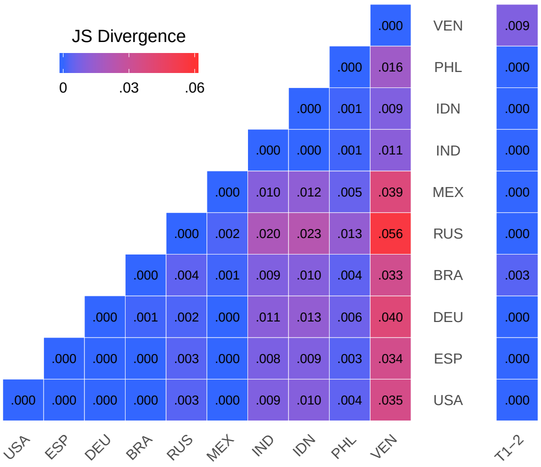


Figure 18. Dependency on CrowdFlower Income JS. This figure shows the JS divergences between the answer distributions of the different countries. The bar on the right shows the JS divergence between `T1` *and* `T2` *for each individual country.*

income group. In the low income group as well as in Venezuela, the proportions were significantly higher. Figure 17 shows the answer distribution of each country.

In terms of distribution differences between countries, Venezuela had the highest JS divergences with other countries, especially with the countries in the high and middle income group. The countries in the high and middle income categories were very similar among each other. Figure 18 shows the JS divergences of the answer distributions.

The reliance of workers on CrowdFlower as a main source of income was mostly stable between T1 and T2, with the exception of Venezuela and, to a lesser extent, Brazil. In Venezuela, consistent with the increase of weekly time spent on the platform, the percentage of workers relying on CrowdFlower as a primary source of income significantly increased from T1 (29%) to T2 (41.5%). In Brazil the percentage was also higher in T2 (15.9%) than in T1 (11.1%).

5.3. Use of Micro-Task Income

The question regarding workers' use of the income earned through micro tasks offered seven answer options (see Table 3) and workers could select one or more of the options. Figure 19 shows the proportion of workers who selected the different expenditure categories, for each country.[18]

In seven out of ten countries, the proportion of workers who reported spending micro-task income on basic expenses such as food, rent, sanitary items or medical care exceeded 40%. The countries with the highest proportion of workers who spent the money on basic expenses were the Philippines and Venezuela. Germany was the country with the lowest percentage of workers spending the money for basic expenses, followed by Spain and Russia. In the USA, despite being a high income country, over 40% of workers reported spending the money on basic expenses.

The three countries in the high income group and Brazil had the highest percentage of workers who stated spending the money on leisure activities such as hobbies, games, holidays or sports. In all other countries except Venezuela, the proportion of workers who reported spending micro-task income on leisure activities was also higher than 30%. In Venezuela, the proportion of workers who reported spending micro-task income on leisure activities was by far lowest of all countries.

A high percentage of crowdworkers indicated that they save or invest the money earned on CrowdFlower, especially in lower income countries. The countries where the highest percentage of workers who chose this response were Venezuela, the Philippines and Indonesia. The USA and Russia had the lowest proportions of workers who reported saving or investing the income from micro tasks.

The USA, Russia and India had the highest proportion of workers who reported spending the money on gifts, while the lowest proportion for this expenditure category was in Venezuela. A moderate percentage of workers stated using the micro-task income for financing their education. This expenditure category was highest in Venezuela, followed by India, Mexico and Indonesia. In most countries, very few workers donate their income from micro tasks to charities, with the exception

[18] Note that the sum of the different categories may be higher than 100% for each country, as workers could choose more than one expenditure category.

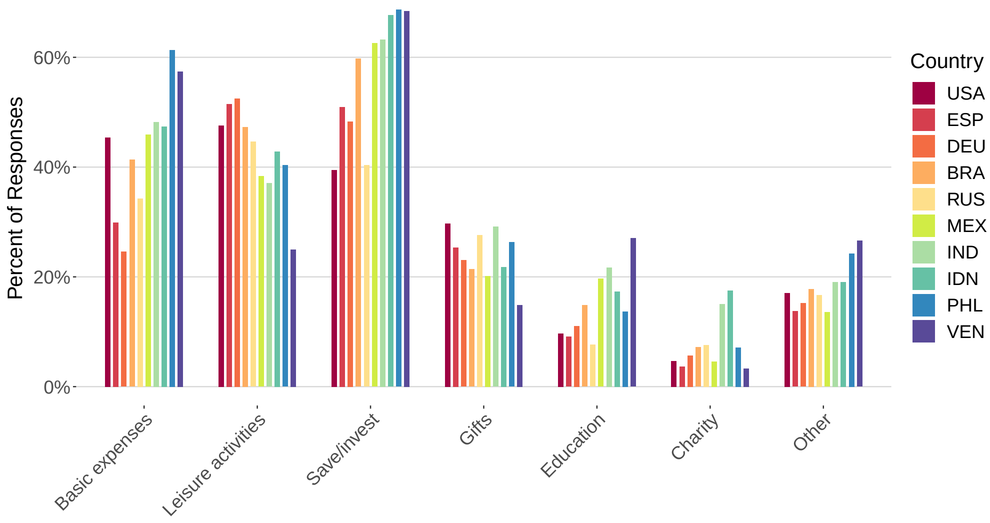
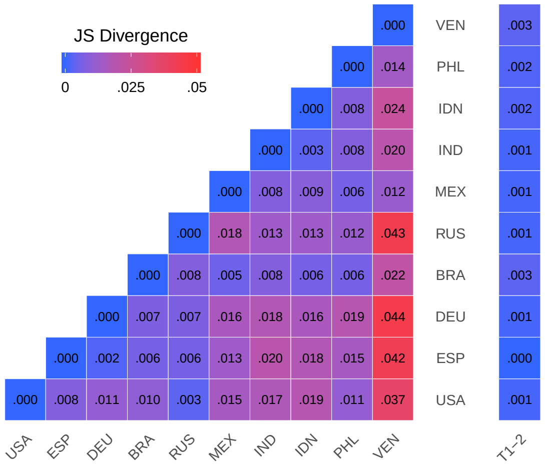


Figure 19. Use of CrowdFlower Income. This figure shows how workers spend their income from micro tasks in the different countries. The bar heights represent the averages of T1 and T2.



	USA	ESP	DEU	BRA	RUS	MEX	IND	IDN	PHL	VEN	T1~2
VEN										.000	.003
PHL									.000	.014	.002
IDN								.000	.008	.024	.002
IND							.000	.003	.008	.020	.001
MEX						.000	.008	.009	.006	.012	.001
RUS					.000	.018	.013	.013	.012	.043	.001
BRA				.000	.008	.005	.008	.006	.006	.022	.003
DEU			.000	.007	.007	.016	.018	.016	.019	.044	.001
ESP		.000	.002	.006	.006	.013	.020	.018	.015	.042	.000
USA	.000	.008	.011	.010	.003	.015	.017	.019	.011	.037	.001

Figure 20. Use of CrowdFlower income JS. This figure shows the JS divergences between the answer distributions of the different countries. The bar on the right shows the JS divergence between T1 and T2 for each individual country.

of India and Indonesia. A significant proportion of workers also stated that they used the money for purposes other than the given categories, especially in the Philippines and in Venezuela.

Figure 20 shows the JS divergences of the answer distributions. As this survey question allowed for multiple answers, we normalized the distributions to sum to one before calculating the JS divergences. We found the largest differences in distribution between Venezuela and the three countries in the high income group as well as Russia. Generally, the countries in the high income group as well as Russia were somewhat similar among each other, and more dissimilar to the countries in the low income group and Venezuela.

Regarding the difference between the time points, Venezuela showed the largest changes. These changes were mostly in the categories *basic expenses* and *education*. While in T1 the country with the highest proportion of workers spending the micro task money for basic expenses was the Philippines, in T2 it was Venezuela. The proportion of Venezuelan workers who reported using the micro-task income money use for basic expenses rose from 52.4% in T1 to 62.4% in T2. This is consistent with Venezuelan workers' increase in relying on CrowdFlower as a primary source of income, as well as their increase in weekly time spent on the platform. The proportion of Venezuelan workers using the money for their education rose from 22% at T1 to 32% at T2.

The second-largest change[19] from T1 to T2 was observed in Brazil. In T2, less Brazilian workers indicated saving or investing their micro-task income, while more Brazilian workers reported spending it on basic expenses and leisure activities. In T2, there was also a lower percentage of workers who reported donating micro-task income to charity than in T1, in all countries of the low income group.

6. CONCLUSION

The work presented in this paper constitutes the first large scale comparison of crowdworker characteristics at the country level that goes beyond an analysis of the two countries that constitute the majority of workers on MTurk. By shedding light on the country-specific differences of the international crowd workforce, this study complements existing research and contributes to a better understanding of this emerging form of work.

We presented an analysis of the demographic composition of the crowd workforce in ten countries and the centrality of micro-task income in workers' lives. We based our analysis on two large samples of crowdworkers from ten different countries, collected at two different points in time on the platform CrowdFlower. Our results reveal significant differences in demographic composition, time spent on the platform, reliance on micro-task income as well as use of micro-task income between the different countries. Furthermore, our results show that the characteristics of the workforce in different countries remained, in most cases, largely stable between the two samples collected eight months apart. While there were changes in the answer distributions of certain characteristics in some countries, the average differences between the countries were larger than the average change over time. These results constitute an important step towards a more comprehensive characterization of the international crowd workforce.

[19] Brazil had a slightly lower JS (0.0025) than Venezuela (0.0028).

Our study has several limitations. While we took great care to account for fluctuations in worker composition (e.g. by the hour of the day or the day of the week) by dividing the starting times of our tasks into different categories, further research on the stability of the different characteristics is needed. Furthermore, due to the nature of micro tasks, our samples are necessarily self-selected. Our samples therefore do not include workers who, for example, exclusively work on repeatable tasks and never accept survey tasks. Lastly, our sample focuses on workers who have sufficient English skills to understand the survey questions. However, this is likely true for the majority of the micro-task workforce on this platform, as workers are expected to understand instructions in English[20] and demand for crowdworkers is driven by Anglophone countries (Kuek et al., 2015).

In future work, we plan to analyze the relationship between demographic characteristics and motivational profiles of crowdworkers, using the Multidimensional Crowdworker Motivation Scale (Posch et al., 2017). Furthermore, future research will be able to use the data presented in this study in order to compare the demographic composition of the crowd workforce with the composition of the general population, and the general workforce, in different countries. Finally, future research focusing on the examination of factors that cause the differences in crowd workforce composition between countries and over time will further contribute to a better understanding of the phenomenon of crowdwork. This paper is relevant for researchers and practitioners interested in the composition of the international crowd workforce.

7. REFERENCES

Amazon Mechanical Turk, . (2016). Amazon Mechanical Turk: Worker Web Site FAQs. http://www.mturk.com/mturk/help?helpPage=worker#how_paid. (2016). Accessed: 2016-12-20.

Berg, J. (2016). Income security in the on-demand economy: findings and policy lessons from a survey of crowdworkers. (2016).

Berinsky, A. J, Huber, G. A, and Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon. com's Mechanical Turk. *Political Analysis* 20, 3 (2012), 351–368.

Bloomberg News, . (2016). Venezuela's Currency Is Collapsing on the Black Market Again. http://www.bloomberg.com/news/articles/2016-11-01/venezuela-s-currency-is-collapsing-on-the-black-market-again. (2016). Accessed: 2016-12-20.

Brabham, D. C. (2010). MOVING THE CROWD AT THREADLESS. *Information, Communication & Society* 13, 8 (2010), 1122–1145. DOI:http://dx.doi.org/10.1080/13691181003624090

Brewer, R, Morris, M. R, and Piper, A. M. (2016). Why would anybody do this?: Understanding older adults' motivations and challenges in crowd work. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 2246–2257.

Buhrmester, M, Kwang, T, and Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6, 1 (2011), 3–5.

Destatis, . (2013). Einkommens- und Verbrauchsstichprobe: Einnahmen und Ausgaben privater Haushalte. *Fachserie 15 des Statistischen Bundesamtes* 4 (2013).

Difallah, D, Filatova, E, and Ipeirotis, P. (2018). Demographics and dynamics of mechanical turk workers. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, 135–143.

European Agency for Health and Safety at Work, . (2015). The future of work: crowdsourcing. https://osha.europa.eu/en/tools-and-publications/publications/future-work-crowdsourcing/view. (2015). Accessed: 2018-10-20.

Goodman, J. K and Paolacci, G. (2017). Crowdsourcing consumer research. *Journal of Consumer Research* 44, 1 (2017), 196–210.

Hirth, M, Hoßfeld, T, and Tran-Gia, P. (2011). Human cloud as emerging Internet application-anatomy of the microworkers crowdsourcing platform. *University of Wurzburg Institute of Computer Science Research Report Series* (2011).

Huff, C and Tingley, D. (2015). "Who are these people?" Evaluating the demographic characteristics and political preferences of MTurk survey respondents. *Research & Politics* 2, 3 (2015), 2053168015604648.

[20] The platform's interface is available exclusively in English.

Ipeirotis, P. G. (2010)a. Analyzing the Amazon Mechanical Turk marketplace. *ACM Crossroads* 17, 2 (2010), 16–21. DOI:http://dx.doi.org/10.1145/1869086.1869094

Ipeirotis, P. G. (2010)b. Demographics of mechanical turk. *CeDER Working Papers* (2010).

Kazai, G, Kamps, J, and Milic-Frayling, N. (2012). The face of quality in crowdsourcing relevance labels: Demographics, personality and labeling accuracy. In *Proceedings of the 21st ACM international conference on Information and knowledge management*. ACM, 2583–2586.

Kuek, S. C, Paradi-Guilford, C, Fayomi, T, Imaizumi, S, Ipeirotis, P, Pina, P, Singh, M, and others, . (2015). *The global opportunity in online outsourcing*. Technical Report. The World Bank.

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information theory* 37, 1 (1991), 145–151.

Martin, D, Carpendale, S, Gupta, N, Hoßfeld, T, Naderi, B, Redi, J, Siahaan, E, and Wechsung, I. (2017). Understanding the Crowd: Ethical and Practical Matters in the Academic Use of Crowdsourcing. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Springer, 27–69.

Naderi, B. (2018). Who are the Crowdworkers? In *Motivation of Workers on Microtask Crowdsourcing Platforms*. Springer, 17–27.

Paolacci, G and Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a participant pool. *Current Directions in Psychological Science* 23, 3 (2014), 184–188.

Paolacci, G, Chandler, J, and Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgment and Decision making* 5, 5 (2010), 411–419.

Pavlick, E, Post, M, Irvine, A, Kachaev, D, and Callison-Burch, C. (2014). The language demographics of amazon mechanical turk. *Transactions of the Association for Computational Linguistics* 2 (2014), 79–92.

Peer, E, Brandimarte, L, Samat, S, and Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology* 70 (2017), 153–163.

Posch, L, Bleier, A, and Strohmaier, M. (2017). Measuring Motivations of Crowdworkers: The Multidimensional Crowdworker Motivation Scale. *CoRR* https://arxiv.org/abs/1702.01661 (2017).

Ross, J, Irani, L, Silberman, M. S, Zaldivar, A, and Tomlinson, B. (2010). Who are the crowdworkers?: shifting demographics in mechanical turk. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Extended Abstracts Volume, Atlanta, Georgia, USA, April 10-15, 2010*, Elizabeth D. Mynatt, Don Schoner, Geraldine Fitzpatrick, Scott E. Hudson, W. Keith Edwards, and Tom Rodden (Eds.). ACM, 2863–2872. DOI:http://dx.doi.org/10.1145/1753846.1753873

Ross, J, Zaldivar, A, Irani, L, and Tomlinson, B. (2009). Who are the turkers? worker demographics in amazon mechanical turk. *Department of Informatics, University of California, Irvine, USA, Tech. Rep* (2009).

Shapiro, D. N, Chandler, J, and Mueller, P. A. (2013). Using Mechanical Turk to study clinical populations. *Clinical Psychological Science* 1, 2 (2013), 213–220.

United Nations Department Of Economic And Social Affairs - Statistics Division, . (2000). Classifications of expenditure according to purpose. https://unstats.un.org/unsd/publication/SeriesM/SeriesM_84E.pdf. (2000). Accessed: 2018-12-02.

U.S. Bureau of Labor Statistics, . (2018). Consumer Expenditure Surveys - CE Survey Materials. https://www.bls.gov/cex/csxsurveyforms.htm. (2018). Accessed: 2018-12-02.

Weinberg, J. D, Freese, J, and McElhattan, D. (2014). Comparing Data Characteristics and Results of an Online Factorial Survey between a Population-based and a Crowdsource-recruited Sample. *Sociological Science* 1 (2014).

3.3.2 Measuring Motivations of Crowdworkers: The Multidimensional Crowdworker Motivation Scale

This article presents the development and validation of the Multidimensional Crowdworker Motivation Scale (MCMS). Furthermore, this article presents the first cross-country and cross-income group comparison of crowdworker motivations, providing answers to the third overarching research question *RQ3* (presented in Section 1.4).

In the article presented in this section, we first set out to address research question *RQ3.1* by aiming to provide a valid measurement instrument for motivations in the microtask context. To that end, we first analyzed the suitability of existing work motivation scales for the context of microtasks. Using data collected on the platform Figure Eight, we evaluated five different models based on two widely-used work motivation scales developed for the traditional work context. Our results showed that the evaluated work motivation scales do not work well within the context of microtasks, but that, nevertheless, both scales contain items that are potentially useful for measuring the motivations of crowdworkers.

Based on the results of this analysis, we set out to develop a model for measuring motivations in the microtask context. We first developed an item pool containing potentially useful items and conducted three rounds of data collection in three different countries to refine the pool and arrive at the final model for the MCMS.

With the final version of the MCMS, we collected data from ten countries selected from three different World Bank income groups for evaluation. An evaluation of the factorial structure of the hypothesized model showed that the model fit the data well. To address *RQ3.2*, we further conducted separate evaluations of the factorial structure in each country and income group, showing that the model had adequate fit in all groups. We then conducted further analyses of the model's validity, including a study on the microtask platform MTurk. The results of these analyses indicated that the MCMS is a reliable and valid measurement of crowdworker motivations within the framework of self-determination theory.

We then set out to address *RQ3.3* by analyzing the cross-country and cross-income group comparability of results obtained with the MCMS. By conducting measurement invariance tests, we demonstrated that the results of the MCMS are comparable across the workforces in different countries and income groups, using the model-estimated latent means of the different motivational dimensions.

Finally, to provide answers to the overarching research question *RQ3*, the article reports on the motivations of crowdworkers in the ten countries included in our study. Our results showed that significant differences exist in the motivational profiles of the microtask workforces of different countries and income groups. However, monetary rewards were the most important motivation for crowdworkers in all countries.

Measuring Motivations of Crowdworkers: The Multidimensional Crowdworker Motivation Scale

LISA POSCH, GESIS–Leibniz Institute for the Social Sciences, Germany and Graz University of Technology, Austria
ARNIM BLEIER and CLEMENS M. LECHNER, GESIS–Leibniz Institute for the Social Sciences, Germany
DANIEL DANNER, University of Applied Labour Studies, Germany
FABIAN FLÖCK, GESIS–Leibniz Institute for the Social Sciences, Germany
MARKUS STROHMAIER, RWTH Aachen University, Germany and GESIS – Leibniz Institute for the Social Sciences, Germany

Crowd employment is a new form of short-term and flexible employment that has emerged during the past decade. To understand this new form of employment, it is crucial to illuminate the underlying motivations of the workforce involved in it. This article introduces the Multidimensional Crowdworker Motivation Scale (MCMS), a scale for measuring the motivation of crowdworkers on microtask platforms. The MCMS is theoretically grounded in self-determination theory and tailored specifically to the context of paid crowdsourced microlabor. The scale measures the motivation of crowdworkers along six motivational dimensions, ranging from amotivation to intrinsic motivation. We validated the MCMS on data collected in ten countries and three income groups. Factor analyses demonstrated that the MCMS's six dimensions showed good model fit, validity, and reliability. Furthermore, our measurement invariance tests showed that motivations measured with the MCMS are comparable across countries and income groups, and we present a first cross-country comparison of crowdworker motivations. This work constitutes an important first step toward understanding the motivations of the international crowd workforce.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **General and reference** → **Surveys and overviews**; • **Social and professional topics** → *Employment issues*; Geographic characteristics;

Additional Key Words and Phrases: Crowdsourcing, crowdworkers, motivation, self-determination theory, scale, validation, invariance

Authors' addresses: L. Posch, GESIS–Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne, 50667, Germany, Graz University of Technology, Inffeldgasse 16c, Graz, 8010, Austria; email: lisa.posch@gesis.org; A. Bleier and F. Flöck, GESIS–Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne, 50667, Germany; emails: {arnim.bleier, fabian.floeck}@gesis.org; C. M. Lechner, GESIS–Leibniz Institute for the Social Sciences, B2,1, Mannheim, 68159, Germany; email: clemens.lechner@gesis.org; D. Danner, University of Applied Labour Studies, Seckenheimer Landstraße 16, Mannheim, 68163, Germany; email: daniel.danner@arbeitsagentur.de; M. Strohmaier, RWTH Aachen University, Theaterplatz 14, Aachen, 52062, Germany, GESIS–Leibniz Institute for the Social Sciences, Unter Sachsenhausen 6-8, Cologne, 50667, Germany; email: markus.strohmaier@humtec.rwth-aachen.de.

2469-7818/2019/09-ART8
https://doi.org/10.1145/3335081

ACM Reference format:
Lisa Posch, Arnim Bleier, Clemens M. Lechner, Daniel Danner, Fabian Flöck, and Markus Strohmaier. 2019. Measuring Motivations of Crowdworkers: The Multidimensional Crowdworker Motivation Scale. *ACM Trans. Soc. Comput.* 2, 2, Article 8 (September 2019), 34 pages.
https://doi.org/10.1145/3335081

1 INTRODUCTION

During the past decade, crowd employment has emerged as a new form of short-term and flexible employment. As such, crowd employment is part of a wider trend in industrial societies toward increasingly flexible work arrangements that are characterized by short-term, market-based contracts [41, 50]. Crowd employment has been defined as a type of employment that "uses an online platform to enable organisations or individuals to access an indefinite and unknown group of other organisations or individuals to solve specific problems or to provide specific services or products in exchange for payment" [59]. While this definition is similar to the concept of crowdsourcing [47], it explicitly includes only those activities that are performed in exchange for payment.

One type of crowd employment platforms are microtask platforms such as Amazon Mechanical Turk[1] (AMT) or CrowdFlower.[2] On microtask platforms, crowdworkers are paid on a per-task basis, and a single task usually pays only a few cents upon completion. The microtasks offered to workers on these platforms are also called "human intelligence tasks" and typically require workers to solve problems that are easy to solve for humans but hard to solve for computers. This characteristic led Amazon Mechanical Turk to coin the term "artificial artificial intelligence" to describe this type of work. Typical microtasks include classification and tagging of text or images, audio and image transcription, and validating addresses of companies on the web. Also more complex tasks such as editing text documents [5], ontology alignment [81] and the evaluation of unsupervised machine learning algorithms (e.g., References [10, 31, 74]) have been successfully deployed on microtask platforms. Anyone, regardless of geographical location or education, can perform microtasks—the only necessary requirement is having access to the Internet.

The emergence of crowd employment and a general trend towards more flexible and shorter-term employment have given rise to policy discussions on social protection and working conditions of crowdworkers (e.g., References [15, 16, 27, 72]). One ongoing discussion is whether crowd employment is to be considered "work" at all, or whether it is mostly considered a spare-time activity by many workers, meaning that remuneration plays only a minor role for them [72]. Estimates of the hourly wage achievable on popular microtask platforms lie between under US$1 and around US$5 [3, 4, 44, 52, 77]. While this amount is above the minimum wage in some countries, in many high-income countries it is far below the wage of any traditional job. Despite this, the rise of crowd employment is an international phenomenon that does not exclude high-income countries. Understanding the underlying motivations of the international, "indefinite and unknown group" of crowdworkers is therefore crucial for understanding this new form of employment.

Although there has been some research on the motivation of crowdworkers, there is currently no theoretically founded scale for comprehensively measuring motivations of crowdworkers in different countries. So far, the question of what motivates people across the world to participate in microtask crowdwork remains largely open.

This article lays the groundwork for understanding the motivations of the international crowd workforce by introducing the Multidimensional Crowdworker Motivation Scale (MCMS) and

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

presenting a case study conducted on a large sample of crowdworkers from ten different countries. The MCMS is theoretically grounded in self-determination theory (SDT) and tailored specifically to the context of paid crowdsourced microlabor. Most items in the MCMS are based on items from existing SDT-based motivation scales developed for the traditional work context, which we adapted to the idiosyncrasies of work on microtask platforms.

The main contributions of this article are (1) an evaluation of two existing SDT-based work motivation scales developed for the traditional work context with respect to their suitability for microtask crowdwork; (2) the development of the Multidimensional Crowdworker Motivation Scale (MCMS) that draws from these existing scales but refines them and adapts them to the context of microtasks; (3) a validation of the MCMS in ten countries and three income groups; (4) an evaluation of the comparability of motivations measured with the MCMS across countries and across income groups; and (5) a first analysis of differences in crowdworker motivations across these countries and income groups. To the best of our knowledge, the MCMS is the first motivation scale developed specifically for the context of crowdsourced microlabor that offers a comprehensive representation of the motivational dimensions according to SDT. Furthermore, it is the first motivation scale for the crowdworking domain that is validated across multiple countries and income groups.

The article is structured in the following way: Section 2 gives a short overview of the different types of motivation as conceptualized by self-determination theory and reviews existing SDT-based work motivation scales. Furthermore, it gives an overview of related work on the motivations of crowdworkers on microtask platforms. In Section 3, we evaluate to what extent SDT-based motivation scales developed for the traditional work context can be successfully applied for measuring crowdworker motivations, and we show the need for a work motivation scale adapted to the idiosyncrasies of the microtask context. Section 4 describes the process of developing the MCMS, and Section 5 presents a validation of the MCMS in ten countries and three income groups. In Section 6, we demonstrate the cross-country and cross-income group comparability of motivations measured with the MCMS. Section 7 presents a first cross-country and cross-income group comparison of crowdworker motivations. Finally, Section 8 concludes this work and discusses the scale's limitations as well as directions for future research.

2 RELATED WORK

Self-Determination Theory and Work Motivation. Self-determination theory (SDT) is a theory of human motivation that was developed by Deci and Ryan [20–22]. The theory specifies three general kinds of motivation that are hypothesized to lie along a continuum of self-determination: *amotivation*, *extrinsic motivation* and *intrinsic motivation*. At the one extreme of the continuum lies *amotivation*, which completely lacks self-determination; at the other extreme lies *intrinsic motivation*, which is completely self-determined [33]. Between these extremes lies *extrinsic motivation*, which is further split up into subtypes with varying degrees of internalisation: *external regulation*, *introjected regulation*, *identified regulation* and *integrated regulation*.

Figure 1 (based on Reference [33]) shows the types of motivation as specified by SDT. *Amotivation* is the absence of motivation, a state of acting passively or not intending to act all. *External regulation* is the least self-determined form of extrinsic motivation. Individuals motivated by external regulation act to obtain rewards or avoid punishments. *Introjected regulation* refers to a form of partially internalized extrinsic motivation that aims at the avoidance of guilt or at attaining feelings of worth [24]. *Identified regulation* is a form of extrinsic motivation with a high degree of perceived autonomy, where the action is in alignment with the individual's personal goals. *Integrated regulation* is the most self-determined form of extrinsic motivation and stems from evaluated identifications that are in alignment with self-endorsed values, goals and needs

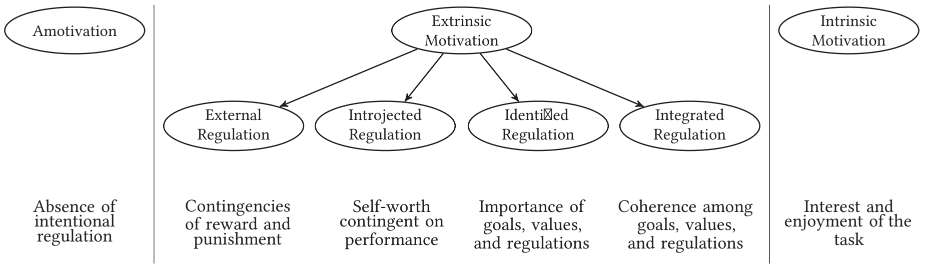

Amotivation
Extrinsic Motivation
Intrinsic Motivation
External Regulation
Introjected Regulation
Identified Regulation
Integrated Regulation
Absence of intentional regulation
Contingencies of reward and punishment
Self-worth contingent on performance
Importance of goals, values, and regulations
Coherence among goals, values, and regulations
Interest and enjoyment of the task

Fig. 1. **Types of motivation.** This figure shows the different types of motivation along the self-determination continuum hypothesized by SDT. The figure is based on Gagné and Deci [33].

[24]. The most self-determined form of motivation is *intrinsic motivation*. This form of motivation is non-instrumental and people act freely, driven by interest and enjoyment inherent in the action [80].

SDT hypothesizes that individuals may internalize an initially external regulation, which then becomes more self-determined. Regulations can be internalized in different ways, depending on the extent to which the individual has integrated it with his or her sense of self [23]. For example, an activity could initially be externally regulated, because it is not perceived as enjoyable by the individual. However, when this activity becomes valued by the person, for example, because it is perceived to be important for his or her personal goals, the regulation for this behavior is internalized (in this case, becoming identified regulation).

While SDT postulates that the different types of internalization fall along a continuum structure, empirical evidence for this continuum hypothesis is inconsistent (e.g., References [11, 37, 46, 58]). For example, Chemolli and Gagné [11] showed that motivations differ more in *kind* than in *degree*, and that SDT-based motivation scales are best represented by multidimensional models. Howard et al. [46] found evidence of a global factor measuring the quantity of self-determination, but they also found that each of the motivation types provided unique information beyond the quantity of self-determination.

Several work motivation scales for the traditional employment context have been developed based on SDT. The first SDT-based work motivation scale was a French scale developed by Blais et al. [6]. Tremblay et al. [88] translated this scale into English and conducted an evaluation in different work environments. The resulting Work Extrinsic and Intrinsic Motivation Scale (WEIMS) measures six factors: amotivation, the four external regulation subtypes and intrinsic motivation. Gagné et al. [34] created the Motivation at Work Scale (MAWS), a scale that measures the four factors external regulation, introjected regulation, identified regulation and intrinsic motivation. The MAWS was validated in French and in English and was partly based on the scale developed by Blais et al. [6].

Later, Gagné et al. [35] developed the Multidimensional Work Motivation Scale (MWMS). The MWMS was validated in seven languages and nine countries and does not include any items from the MAWS. The MWMS measures six first-order factors (amotivation, material external regulation, social external regulation, introjected regulation, identified regulation, and intrinsic motivation) and one second-order factor (external regulation). Work motivation scales such as MWMS, MAWS, and WEIMS investigate motivations at the domain level of analysis, meaning that they measure the general motivation to perform a job as opposed to specific tasks within a job.

Crowdworker Motivation on Microtask Platforms. Compared to work motivation in the traditional employment context, research on motivation in the microtask context is still scarce and scattered. Most research investigating the motivations of workers on microtask platforms has focused on the platform Amazon Mechanical Turk (AMT). Consequently, most studies have focused on American and Indian crowdworkers, which constitute the vast majority of workers on AMT[3] [48, 49, 77]. This country distribution is likely due to the fact that workers can receive money from AMT in the USA and in India while workers from other countries are paid in Amazon.com gift cards [89].

Studies that investigated crowdworker motivation suggest that there are different motivations for participating on microtask platforms. For example, one early study on the reasons crowdworkers have for participating on AMT was conducted by Ipeirotis [49]. In this study, the author asked the multiple-choice question "Why do you complete tasks in Mechanical Turk?", offering six response options. He found that more Indians than Americans treat AMT as a primary source of income, and that few Indian workers report the reason "To kill time." Hossain [45] created a classification of motivation in online platform participation, listing extrinsic and intrinsic motivators and incentives.

Kaufmann et al. [51] developed an early model for measuring crowdworker motivations on AMT, differentiating between enjoyment-based motivation, community-based motivation, immediate payoffs, delayed payoffs, and social motivation. They used a sample composed of Indian and US workers on AMT and found that the construct with the highest score was "immediate payoffs," i.e., payment. Their study further found that the pastime score correlated positively with household income and negatively with the weekly time spent on AMT, and that workers who spend a lot of time on AMT may be motivated differently than workers who spend little time on the platform.

Antin and Shaw [1] used a list experiment to investigate social desirability effects in motivation self-reports of crowdworkers from the USA and India on AMT. Using the four items "to kill time," "to make extra money," "for fun," and "because it gives me a sense of purpose," they found that U.S. workers tended to over-report all four reasons while Indian workers tended to over-report "sense of purpose" and under-report "killing time" and "fun."

For measuring extrinsic motivations of crowdworkers, Naderi et al. [65] evaluated a four-factor model using a subset of WEIMS items on a sample of U.S. workers on AMT. In this model, identified and integrated regulation are merged into one factor, and the intrinsic motivation factor is omitted. After a first version of the present study was published [75], Naderi adapted and extended their scale to include intrinsic motivation [64]. The adapted scale, named the Crowdwork Motivation Scale, measures five constructs, three of which are measured by WEIMS items (amotivation, external regulation and identified regulation). The scale was evaluated on three samples of workers on AMT (N = 170, 90, and 86).

In addition to these few quantitative studies, several qualitative studies on the motivations of crowdworkers have been conducted. For example, Gupta et al. [38, 39] investigated, among other aspects, the motivations of Indian crowdworkers on AMT and Martin et al. [61] studied the content of a forum for AMT users. Other research related to the motivations of crowdworkers includes measuring the impact of motivation on performance [76] and manipulating motivations via task framing [9] or achievement feedback [57].

In sum, these studies demonstrate that there are meaningful differences in crowdworkers' motivations to participate on microtask platforms. However, systematic and theory-driven inquiries into the motivations of crowdworkers remain in short supply. SDT-based work motivation scales may offer a suitable foundation for such inquiries, but the applicability of such scales to the microtask domain has yet to be established.

[3] American and Indian crowdworkers currently constitute over 80% of the worker population on ATM (also see http://demographics.mturk-tracker.com/#/countries/all).

3 SUITABILITY OF EXISTING WORK MOTIVATION SCALES

Crowd employment on microtask platforms is similar to traditional employment in the sense that in both contexts, workers provide a service in exchange for payment. In both contexts, tasks need to be completed, and these tasks are usually specified by the employer/requester and executed by the employee/crowdworker. However, several fundamental aspects of work on microtask platforms differ from the traditional employment context. For example, on microtask platforms, the relationship between the requester and the worker is often completely anonymous and extremely short-lived, often lasting only a few minutes. Furthermore, there is only a minimal amount of communication between the requester and the workers, often not exceeding the static one-way communication via written task instructions. Regarding the social environment, there is often no communication or collaboration between co-crowdworkers, as microtasks are intended to be completed as individual work.

To determine to what extent existing SDT-based work motivation scales that were developed for the traditional work context are suitable for application in the microtask context, we conducted an evaluation of two work motivation scales, the WEIMS [88] and the MWMS [35], with crowdworkers on CrowdFlower.[4] The microtask market is dominated by two platforms, CrowdFlower and AMT, which are estimated to share 80% of all revenue generated in the microtask market, with revenues being approximately equal [55]. Our reason for choosing CrowdFlower over AMT is that we aim to provide a motivation scale suitable for an international comparison of crowdworker motivations, instead of exclusively focusing on crowdworkers based in the USA and in India. We consider CrowdFlower to be better suited for this task as it pays workers via independent partner channels[5] and therefore attracts a more international crowd-workforce.

For the scale stems and items of the WEIMS and the MWMS to be conceptually applicable to the crowdworking domain, we had to make minimal adaptations to the scales before evaluating them in the microtask context. For WEIMS, we changed the stem "Why do you do your work?" to "Why do you do CrowdFlower tasks?"[6] and replaced the word "it" (referring to "your work") in the items with "CrowdFlower tasks." The stem of MWMS "Why do you or would you put efforts into your current job?" was changed to "Why do you or would you put efforts into CrowdFlower tasks?" and words in the items referring to "your current job" were replaced with "CrowdFlower tasks." Additionally, one item in the MWMS was conceptually not applicable to the domain and had to be adapted. There is no equivalent to "losing one's job" on microtask platforms. The closest concept on CrowdFlower is failing many quality control questions, which results in a lower worker account accuracy and consequently in less tasks being offered to the worker. Therefore, the item "Because I risk losing my job if I don't put enough effort in it." was changed to "Because I risk not being offered enough tasks if I don't put enough effort into them."

Respondents answered both scales along a 7-point Likert-type scale. We adopted the verbal descriptions of the scale's endpoints from the original scales: For the adapted WEIMS (A-WEIMS), the scale ranged from "does not correspond at all" (1) to "corresponds exactly" (7) and for the adapted MWMS (A-MWMS), the scale ranged from "not at all" (1) to "completely" (7).

For both the A-WEIMS and the A-MWMS, we collected responses from 500 crowdworkers residing in the USA. The surveys were posted as tasks on CrowdFlower and anonymity was ensured

[4] CrowdFlower changed its name to Figure Eight in February 2018 [8].

[5] http://www.crowdflower.com/labor-channels/.

[6] We chose to use the term "CrowdFlower tasks" in the stem questions and in the items instead of a more general term, because workers who are logged into CrowdFlower via the partner channels see that they are doing "CrowdFlower tasks." We can therefore assume that workers know what CrowdFlower tasks are. In contrast, general terms like "microtasks" are widely used in scientific publications and sometimes in the media, but we cannot be sure that workers on CrowdFlower understand this term as it does not appear frequently on partner channel websites or on the platform itself.

in the description of the tasks. After removing spammers (also see Section 5), the sample size was 424 for the A-WEIMS and 414 for the A-MWMS. This entails a subject-to-item ratio higher than 20:1, which is a suitable ratio for factor analysis [32, 70]. The questionnaires adhered to the default design of the CrowdFlower platform and the design was very similar to that used in the later validation of the MCMS scale (see figures in Appendix C).

Confirmatory Factor Analysis. Confirmatory factor analysis (CFA) is a multivariate data analysis technique used to test how well latent constructs, specified according to theory, represent reality according to the data gathered [40]. We used CFA to evaluate the psychometric quality (i.e., the quality of the measurement of the latent constructs) of the A-WEIMS and the A-MWMS. In particular, we evaluated the factor structure of the items with CFAs, using the R packages lavaan [78] and semtools [83].

Because of non-normality of the item distributions, we used a robust maximum likelihood estimator (as suggested in, e.g., References [19, 28]). By specifying a robust maximum likelihood estimator, the model parameters were estimated with robust standard errors, and a Satorra-Bentler (S-B) scaled test statistic is reported [78, 82].

In the context of CFA, the validity of a measurement model is evaluated via a range of goodness-of-fit (GOF) measures. These GOF measures assess the extent to which the theory (as specified in the model) represents reality (i.e., the data). Once acceptable levels of GOF measures are established, other aspects of construct validity can be evaluated [40]. For a further discussion of construct validity, see Section 5. We evaluated the model fit based on the absolute[7] fit measures *root mean squared error of approximation* (RMSEA) and *standardised root mean square residual* (SRMR) as well as the incremental[8] fit measures *comparative fit index* (CFI) and *Tucker-Lewis index* (TLI). In line with current conventions for judging model fit (e.g., References [53, 60]), we chiefly relied on the CFI, RMSEA and and SRMR to assess model fit[9] and judged model fit to be acceptable according to the following criteria: A well-fitting model should have an RMSEA of less than 0.06, a SRMR of less than 0.08, and a CFI and TLI higher than 0.95 (but at least 0.9 to be acceptable). Furthermore, we report the (S-B scaled) Chi-Square test statistic but do not rely on it for determining model fit as it is very sensitive to sample size (e.g., Reference [2]).

Measurement Models. We tested the following models for the adapted WEIMS: (1) The original WEIMS model with six factors (`A-WEIMS-M1`); (2) an alternative five-factor WEIMS model with identified regulation and integrated regulation loading onto a single factor (`A-WEIMS-M2`); and (3) the four-factor, 12-item subset of WEIMS items used by Naderi et al. [65] to measure the extrinsic motivations of workers on Amazon Mechanical Turk (`A-WEIMS-M3`).[10] Our rationale for evaluating the alternative model `A-WEIMS-M2` is that the integrated regulation factor has been shown to be poorly separable from identified regulation and intrinsic motivation (e.g., References [88, 90]), which is also one of the reasons for why the MWMS does not include an integrated regulation factor [35].

For the A-MWMS, we tested the model originally hypothesized by Gagné et al. [35] (`A-MWMS-M1`) and the model that had the best fit in Reference [35] (`A-MWMS-M2`). Furthermore, we tested a six-factor model in which material external and social regulation are separate factors, omitting the second-order external regulation factor. We tested this model with a hypothesized correlation of

[7] Absolute GOF measures measure how well the model represents the data, independently of other, alternative, models [40].

[8] Incremental GOF measures compare the specified model with a baseline model where all variables are uncorrelated [40].

[9] A discussion on the guidelines for determining model fit can be found in Hooper et al. [42].

[10] We did not include the extended and adapted version of the 12-item subset of WEIMS items [64] in the evaluation as it was published a year after the development and validation of the MCMS was completed and made available as a first version [75].

Table 1. Evaluation of Existing Work Motivation Scales

Scale/Model	N	S-Bχ^2	df	CFI	TLI	RMSEA	RMSEA 90% CI		SRMR
A-WEIMS-M2	424	500.58	125	0.908	0.888	0.084	0.077	0.091	0.067
A-WEIMS-M3	424	110.57	48	0.969	0.957	0.055	0.043	0.068	0.043
A-MWMS-M1	414	828.27	143	0.818	0.782	0.108	0.101	0.114	0.170
A-MWMS-M3	414	667.90	139	0.859	0.827	0.096	0.089	0.103	0.155
A-MWMS-M4	414	532.55	137	0.895	0.869	0.084	0.077	0.091	0.087

This table shows the goodness-of-fit measures for the different models based on existing SDT-based work motivation scales that were minimally adapted to the crowdworking domain.

zero between intrinsic motivation and both external regulation factors (A-MWMS-M3) as well as without this correlation restriction (A-MWMS-M4).

Results. Table 1 shows the goodness-of-fit statistics of the different models. None of the evaluated models, with the exception of the four-factor model A-WEIMS-M3, which does not measure intrinsic motivation or social external regulation, had an acceptable model fit on our data.

The A-WEIMS model with six factors (A-WEIMS-M1) could not be estimated due to the factors identified regulation and integrated regulation not being distinguishable from another, which resulted in the covariance matrix of the factors not being positive definite. This is consistent with the findings of Naderi et al. [65]. Also for the alternative five-factor model A-WEIMS-M2, the goodness-of-fit measures were outside the acceptable range.

The four-factor model A-WEIMS-M3 was the only evaluated model with a good fit.[11] However, besides the drawbacks that the model does not measure intrinsic motivation or social external regulation, and two of the four factors are measured by only two items each, it has additional limitations: It includes items for measuring the amotivation construct that were criticized by Gagné et al. [35] for resembling low satisfaction of the need for competence rather than measuring amotivation (e.g., "I ask myself this question, I don't seem to be able to manage the important tasks related to this work."). Besides this criticism regarding content validity (i.e., regarding the extent to which the items correspond to the conceptual definition of the construct [40]), the amotivation construct also had an average variance extracted (AVE) below 0.5 (also see Section 5), suggesting low convergent validity [40].

An examination of the estimated correlations of A-WEIMS-M3 reveals that the second-strongest positive correlation (0.38) is between the amotivation construct and the identified/integrated regulation construct.[12] This questions nomological validity (i.e., whether the relationships between the constructs correspond to theory and prior research [40]) as it is inconsistent with both SDT and the correlation patterns reported by other SDT-based motivation scales, which report a low to moderate negative correlation between these constructs (e.g., References [26, 35, 67, 88]).

Of the evaluated A-MWMS models, the six-factor model with separate factors for social external and material external regulation and no correlation restrictions had the best fit. However, the fit was still not acceptable, with all goodness-of-fit measures falling outside acceptable ranges. The fit measures for A-MWMS-M2 are not included in Table 1, because the covariance matrix of the factors was not positive definite. In sum, these results suggest that the existing scales do not allow to measure the crowdworker motivation validly. We thus decided to conduct an in-depth item analysis.

[11] Note, however, that this evaluated subset of A-WEIMS is not identical to the one used by Naderi et al. [65] as a result of the adaptations made in item wording. Naderi et al. [65] reported a lower CFI (0.931) and a higher RMSEA (0.088).

[12] Naderi et al. [65] reported an even higher correlation of 0.48.

To identify which items in A-WEIMS and A-MWMS were most problematic in the microtask context, we conducted an exploratory factor analysis (EFA) on both samples. In contrast to CFA, factors in EFA are not specified according to theory, but derived from the data [40]. Therefore, this technique is useful for investigating the underlying structure of a set of items, using the gathered data as a starting point. For all exploratory factor analyses, we used oblique rotation (promax), because—in line with SDT—we expected the factors (i.e., motivational dimensions) to correlate.

For the A-WEIMS, we conducted an EFA with four factors: We removed the amotivation items due to the problems described above and, because of the CFA results, we expected the integrated regulation items and the identified regulation items to load onto a single factor.[13] The most problematic remaining items were "Because this type of work provides me with security," which had the highest loading on the the same factor as the items from the identified/integrated regulation factor and "Because I want to be a 'winner' in life," which did not have a high loading on any factor. This item was also critiziced by Gagné et al. [35] for being culturally sensitive. Removing these items yielded a solution with four factors (material external regulation, introjected regulation, identified/integrated regulation and intrinsic motivation), whereby two of the factors only had two items remaining.

For the A-MWMS, we conducted an EFA with six factors.[14] Again, one of the most problematic items was related to security ("Because it gives me greater job security if I put enough efforts into doing CrowdFlower tasks"), which did not load onto one factor with the other items from the material external regulation factor, but instead had a high loading on the same factor as the items from the identified regulation construct. Also two items from the introjected regulation construct ("Because I have to prove to myself that I can." and "Because it makes me feel proud of myself.") had moderately high loadings (0.57 and 0.58) on this factor. After removing these three items, the introjected regulation construct as well as several other items were still problematic. Iteratively removing all problematic items yielded a solution with amotivation measured by two items, external material regulation measured by only one item, and either no introjected regulation factor or one that was very poorly distinguishable from the social external regulation factor.

Thus, our results show that the evaluated work motivation scales developed for the traditional work context do not work well within the crowdworking context when only minimal adaptations in item wording are made. While both scales contain items that are potentially useful for measuring the motivations of crowdworkers, neither of the scales is an accurate measure of the full spectrum of motivational dimensions in the microtask context. For the development of a reliable motivation scale that fills these gaps and measures the motivations of crowdworkers on all dimensions proposed by SDT, further adaptations are needed.

4 DEVELOPMENT OF THE MCMS

The results of our factor analyses conducted on the slightly modified WEIMS and MWMS underscore the necessity for developing a new scale for measuring the motivations of crowdworkers that is adapted to the idiosyncracies of the crowdwork environment. To meet this necessity, we developed the Multidimensional Crowdworker Motivation Scale (MCMS). The MCMS was developed to provide a psychometrically sound scale that covers the motivational dimensions proposed by self-determination theory and that can be answered by crowdworkers in a limited amount of time.

[13] An EFA with five factors showed that one item from the identified regulation construct and one item from the integrated regulation construct loaded onto a separate factor, while all other items from these constructs loaded onto a single factor. Therefore, we proceeded with the four-factor version.

[14] An EFA with five factors showed that the items from the material and social external regulation constructs did not load onto the same factor.

We proceeded in three steps. First, we compiled a pool of items conceptually suitable for the characteristics of the crowdworking domain. An item pool is a set of candidate items that reflect the latent constructs the scale intends to measure. During the development of a scale, this set of candidate items is then reduced and refined by deleting items that exhibit undesirable properties such as high loadings on multiple factors or no high loadings on any factor, with the goal of arriving at a final scale with optimized reliability and scale length (e.g., References [25, 93]). Second, we thus selected items from the pool based on an exploratory factor analysis on a sample of workers residing in the USA. Third, we further reduced and refined the item pool based on exploratory factor analyses on samples from Spanish[15] and Indian crowdworkers.

Item Pool Generation. The results of our evaluation described in the previous section indicate that both the A-WEIMS and the A-MWMS contain items that are potentially useful for application in the microtask context. Therefore, for compiling the item pool, we first included all items from the A-WEIMS[16] and the A-MWMS that were not among the most problematic items according to the results of the factor analyses. We retained the adaptations to item wording. To extend the pool, we added semantically suitable items from the SDT-based motivation scales developed in References [67] and [26] as well as nine new items developed by the authors.[17] The total number of items in the pool was 44.

To ensure content validity (i.e., the extent to which the items correspond to the theoretical constructs [40]), we based the majority of candidate items on existing scales, and we closely followed the definitions of the constructs during the selection of candidate items from existing scales as well as during the creation of the new items. We used the construct definitions provided in the Handbook of Self-Determination Theory Research [23] and a publication by Gagné and Deci discussing the constructs in the context of work motivation [33].

We used the stem phrasing "Why do you or would you put efforts into doing CrowdFlower tasks?", adapted from MWMS, to capture both actual and latent motivations. The items were rated along a 7-point Likert-type scale ranging from "not at all" (1) to "completely" (7).[18]

First Round of Data Collection. For a first selection of items from the pool, we collected answers from 1,000 crowdworkers residing in the USA and conducted an exploratory factor analysis on their responses. Consistent with the findings from Section 3, we found that the items intended to measure material external regulation and social external regulation did not load on the same factor. Therefore, we aimed at a six-factor model with the two separate external regulation factors (social and material). Items that had insufficient loadings on the appropriate factor (<0.5), items that loaded on a factor other than the hypothesized one, and items with high cross-loadings (>0.35) were iteratively removed from the initial pool, creating a reduced item pool with 36 items.

Second Round of Data Collection. We conducted a second round of data collection with this reduced item pool, collecting responses from 1,000 Spanish and 1,200 Indian crowdworkers.[19] The additional 200 responses from the Indian crowdworkers were requested because of the high

[15] In this article, we use country demonyms synonymously with the location of workers for better readability.

[16] Notably, this also included all items from `A-WEIMS-M3` except for the amotivation items due to the problems described in Section 3.

[17] The development of the new items focused mainly on the material external regulation construct, as material external regulation is a construct present only in scales intended to measure motivation in a context where material rewards (such as money) are relevant. It is therefore not present in most SDT-based motivation scales.

[18] The verbal descriptions for each scale point were adopted from the MWMS [35] and shown to the participants in the task instructions: 1 = "not at all," 2 = "very little," 3 = "a little," 4 = "moderately," 5 = "strongly," 6 = "very strongly," 7 = "completely."

[19] India was selected, because it has a significant population of crowdworkers on different platforms. Spain was selected to include a European country with a significant population of crowdworkers.

amount of spam received in this group (also see Section 5), with the aim of achieving an item-to-response ratio of close to 1:20. Again, we iteratively removed items with low loadings (with the higher threshold of <0.7 if more than three items were left for this construct) or high cross-loadings (with a threshold of >0.3). Furthermore, if two items were phrased very similarly and the factor had more than three items remaining, we removed the item with the lower loading.

The final MCMS contains 18 items, with three items measuring each factor. Of the 18 final items, five items (Am2, Introj2, Introj3, Intrin1, Intrin3) were adapted from Reference [35], four items (ExMat2, Ident1, Ident2, Ident3) from Reference [88], two items (Am3, Introj1) from Reference [26], two items (ExSoc2, Intrin2) from Reference [67], and five items (Am1, ExMat1, ExMat3, ExSoc1, ExSoc3) are new (but semantically based on items from existing scales). Like other SDT-based work motivation scales such as the MWMS [35], the MCMS aims to measure motivations for putting effort into the job (in this case microtasks) in general, as opposed to measuring the motivations for specific tasks within a job. Table 14 in Appendix A shows the scale.

5 VALIDATION OF THE MCMS

Data Collection. With the final 18-item version of the MCMS, we collected data from 10 countries, with 900 participants from each country, for validation. We selected countries from three World Bank income groups[20]: high-income, upper-middle-income, and lower-middle-income. From each of the three income groups, we selected three countries with high activity on CrowdFlower. The countries were selected according to the following criteria: First, the country had to be active on CrowdFlower (either high in the Alexa[21] ranking or one of the top contributing countries in at least one of the partner channels). Second, we aimed for a high cultural diversity overall as well as within the income groups. For the high-income group, the selected countries were USA, Germany and Spain. The upper middle-income group contains Brazil, Russia and Mexico, and the lower middle-income group is comprised of India, Indonesia and the Philippines. Note that in the rest of this article, we use the group label "Middle Income" (MID) for the upper middle-income group and "Low Income" (LOW) for the lower-middle-income group for better readability.

In addition, we collected responses from Venezuela, because it was the most active country on CrowdFlower at the time of data collection, with CrowdFlower receiving 18.5% of traffic from this country.[22] However, we did not include Venezuela in the data grouped by income, because we believe it represents a special case: At the time of data collection, the US$ earned on CrowdFlower could be sold on the black market at a rate several orders of magnitude higher than the official exchange rate [66].

To capture a diverse sample of crowdworkers in each country, the starting times of the survey were divided into three groups: (1) 300 responses were requested during typical working hours (8:00 am to 5:00 pm in the appropriate time zone), (2) 300 responses were requested in the evening (6:00 pm to 11:00 pm in the appropriate time zone) and finally, (3) 300 responses were requested during weekends. We made the survey available to workers of all CrowdFlower levels. The data was collected in October and November 2016. A full description of the demographic characteristics of our sample can be found in Posch et al. [73] and on our website [29].

Task Interface and Payment. The items in the MCMS were randomly permuted and presented to crowdworkers as a task on CrowdFlower. Besides the MCMS items, the task also included a section with demographic questions and questions about money use, as well as a section in which workers were instructed to think of five reasons for why they do tasks on CrowdFlower and asked

[20] http://databank.worldbank.org/data/download/site-content/CLASS.xls.
[21] http://www.alexa.com/.
[22] Data obtained from http://www.alexa.com/.

to write down these reasons. Anonymity was guaranteed in the task instructions. The interface[23] of the task is shown in Appendix C. After completing a task, crowdworkers on CrowdFlower are asked by the platform to judge the task according to different criteria, one of them being the clarity of the task instructions and interface. The question asked is *"How clear were the task instructions and interface?"* and workers are asked to answer on a five-point scale ranging from "very unclear" to "very clear." The average of the workers' responses was higher than 4.0 in all countries except for Brazil (3.6) and Indonesia (3.8). This indicates that the task instructions and interface was perceived as clear in most countries, and as "somewhat clear" in Brazil and Indonesia.

We aimed for a payment similar to most other tasks on CrowdFlower to minimize population bias in the responses. Based on studies reporting average earnings on microtask platforms (e.g., References [3, 44, 52]) and on the first author's experience as a crowdworker on CrowdFlower, we payed US$0.1 for the task excluding platform fees. The survey presented by the platform to workers after completing a task includes a question about task payment (*"How would you rate the pay for this task relative to other tasks you've completed?"*) that crowdworkers answer on a five-point scale ranging from "much worse" to "much better." For the different countries, the averages of the workers' responses ranged from 3.5 (in Germany) to 4.1 (in Mexico and India), indicating that our task payment was equal to or "somewhat better" than other tasks according to the workers' perception.

Spam Detection. We expected a significant amount of spam in the responses, such as people not reading the questions and clicking randomly or workers accepting the task despite having insufficient English skills. To counteract a high amount of noise in the dataset, we included three test items in the motivation scale section of the CrowdFlower task, and an additional test question in the demographics section.[24] Employing test questions is a common method to implement attention checks in survey design (e.g., References [54, 69, 92]) and has also been employed in crowdsourcing research tasks (e.g., Reference [71]).

The three test items in the motivation scale section of the task asked participants to answer with a specific ranking on a 7-point scale, and the test question in the demographics section consisted of the question "Are you paying attention to the questions?" with the possible answers "No," "Yes," and "I don't know" selectable from a drop-down list. These questions ensured that less than 0.1% $((\frac{1}{7})^3 * \frac{1}{3})$ of spammers passed the test questions, assuming that all four questions were answered at random. Table 2 shows the percentage of spam and the sample size after spam removal in each country and income group. The table also introduces the country and group codes used in the remainder of this article.

Hypothesized Model. Our hypothesized model measures six constructs and is depicted in Figure 2. The inclusion of a social external regulation construct in addition to material external regulation was adopted from the MWMS [35], because both social and material rewards are important in the work context [35, 85]. As suggested by our evaluation of existing scales and our results from exploratory factor analysis on the MCMS item pool, we modeled material external and social external regulation as two separate factors. Hence, our hypothesized model is a six-factor model in which all factors are first-order factors. Note that the material and social external regulation are not adjacent factors in the continuum hypothesized by SDT but occupy the same spot.

[23] English was chosen as the interface language for all countries for two reasons: First, CrowdFlower's default interface language is English and all workers are expected by the platform to understand instructions in English. This is underscored by the fact that "English" is not selectable in the requester interface when choosing crowdworkers of a specific language. Second, translating stems and items has a risk of introducing semantic mismatches. We hence weighted the risk of introducing translation mismatches higher than potential errors due to non-native speakers' misinterpretations.

[24] As CrowdFlower does not offer built-in quality control mechanisms for survey-type tasks, we did not use any platform-specific quality control mechanisms.

Table 2. Sample Sizes and Percentage of Spam Received

Group	Code	N_{raw}	Spam	N_{clean}
All	ALL	9,000	35 %	5,857
High Income	HIGH	2,700	28 %	1,952
Middle Income	MID	2,700	32 %	1,834
Low Income	LOW	2,700	44 %	1,508
USA	USA	900	20 %	721
Spain	ESP	900	25 %	677
Germany	DEU	900	38 %	554
Brazil	BRA	900	45 %	496
Russia	RUS	900	25 %	677
Mexico	MEX	900	27 %	661
India	IND	900	32 %	608
Indonesia	IDN	900	55 %	401
Philippines	PHL	900	45 %	499
Venezuela	VEN	900	37 %	563

This table shows the sample sizes of the different groups before and after spam removal, as well as the percentage of spam received.

Whereas SDT hypothesizes that intrinsic motivation does not correlate with external regulation [35, 79], Chemolli and Gagné [11] found that intrinsic motivation correlates with external regulation significantly for both the MWMS and the Academic Motivation Scale [90]. Based on these findings, we decided not to restrict these correlations to zero in our hypothesized model but to evaluate the model fit of both the correlation-restricted and the unrestricted model. No cross-loadings were hypothesized.

Descriptive Statistics and Internal Consistency. Table 3 summarizes the observed factor means as well as standard deviations for each of the countries and income groups in our data. Table 4 displays the Pearson correlations of the observed factor means. We used Cronbach's alpha statistic [17] to assess the internal consistency of the MCMS. Table 5 displays the values of alpha for each country and income group. Alpha provides an estimate of the lower bound of internal consistency. As a rule of thumb, values above 0.7 are considered acceptable, values between 0.6 and 0.7 questionable, values between 0.5 and 0.6 poor, and values below 0.5 unacceptable [36]. In most countries and groups, alpha exceeded 0.7 for each construct. Exceptions to this were the amotivation factor in Brazil and Venezuela as well as the material external regulation factor in Brazil and Indonesia with values between 0.5 and 0.7. Therefore, when interpreting results from these factors and groups, care should be taken. In addition to Cronbach's alpha, we calculated McDonald's coefficient omega [62] for assessing the reliability of the MCMS. The values for coefficient omega are shown in Appendix B. Compared to Cronbach's alpha, the values of coefficient omega are equal or slightly higher.

Confirmatory Factor Analysis. To validate the factor structure of our hypothesized model, we conducted a confirmatory factor analysis. Table 6 shows the results of the confirmatory factor analysis of the hypothesized model. The first item of each factor in Table 14 served as the marker variable (i.e., the loading of this item was fixed to one). The analysis was conducted on the entire dataset as well as on each group separately. As in Section 3, we followed current conventions for evaluating model fit (e.g., References [42, 53, 60]).

The results show that the hypothesized model had adequate fit overall as well as in all groups. The CFI was above 0.95 in all groups except Indonesia and Mexico (with 0.931 and 0.948, respectively). RMSEA was lower than 0.06 in all groups and SRMR was lower than or equal to 0.05 in all

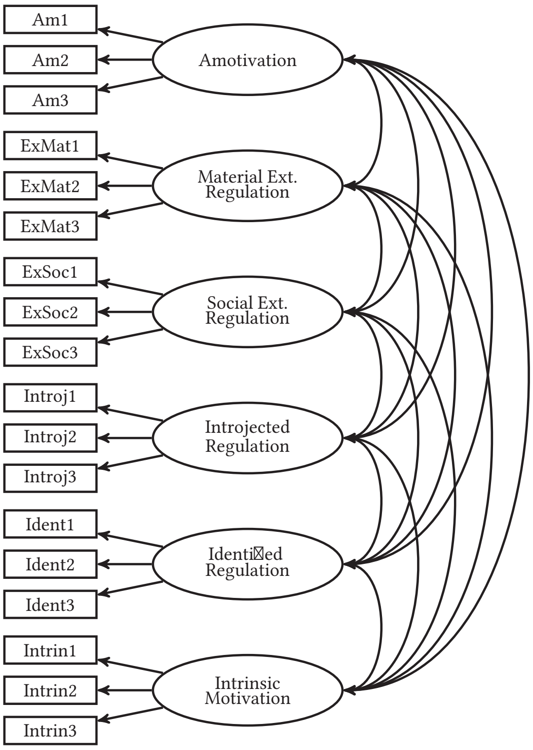


Fig. 2. **Factor structure of the MCMS.** This figure shows the factor structure of the hypothesized MCMS model. Items (i.e., measured variables) are represented by rectangles, ovals represent latent constructs, curved arrows indicate a correlational relationship between latent constructs and straight arrows indicate the dependence relationship between latent constructs and items. The individual items are shown in Appendix A.

groups. We consider the fit measures for Mexico and Indonesia to be marginally acceptable, but some care should be taken when interpreting the results from these countries. Table 7 shows the item loadings and intercepts estimated by the hypothesized model fitted to the entire sample and Table 9 shows the estimated correlations between the constructs. All estimated factor correlations are statistically significant at $p < 0.01$ (most at $p < 0.001$).

The alternative model, which restricts the correlations of the external regulation factors (material external regulation and social external regulation) with intrinsic motivation to zero, did not have an acceptable fit: While CFI was close to 0.95 for most groups (0.948 for the total sample), SRMR was high for the total sample (0.094) as well as for all other groups, ranging between 0.074 in the USA and 0.113 in Russia.

Compared to other SDT-based approaches to measuring crowdworker motivations [64, 65], the MCMS achieved a better fit while measuring additional constructs. Other SDT-based scales used to measure motivations in the microtask context [64, 65] did not report a well-fitting model for sample sizes larger than 100 due to a high RMSEA (>0.08 for the 12-item-subset of WEIMS [65] and >0.09 for the Crowdwork Motivation Scale [64]). For the Crowdwork Motivation Scale, the fit measures were better on two smaller samples (N = 90 and N = 86), but RMSEA was still high (>0.072 for

Table 3. Manifest Scale Means and Standard Deviations

Group	Amotivation	Material	Social	Introjected	Identified	Intrinsic
ALL	1.84 (1.09)	6.05 (1.08)	2.47 (1.58)	2.25 (1.46)	4.27 (1.73)	5.67 (1.26)
HIGH	2.06 (1.21)	5.86 (1.18)	1.99 (1.37)	1.91 (1.29)	3.56 (1.73)	5.27 (1.37)
MID	1.86 (1.06)	6.12 (1.02)	2.68 (1.60)	2.59 (1.58)	4.53 (1.65)	5.82 (1.18)
LOW	1.72 (1.02)	6.13 (0.99)	2.74 (1.68)	2.31 (1.44)	4.60 (1.57)	5.90 (1.11)
USA	1.91 (1.17)	5.86 (1.25)	1.67 (1.19)	1.58 (1.12)	3.38 (1.74)	5.28 (1.43)
ESP	2.10 (1.20)	5.96 (1.07)	2.46 (1.48)	2.34 (1.45)	3.89 (1.71)	5.37 (1.33)
DEU	2.23 (1.26)	5.72 (1.21)	1.82 (1.31)	1.83 (1.15)	3.40 (1.67)	5.15 (1.35)
BRA	1.89 (0.97)	6.28 (0.87)	2.59 (1.60)	2.38 (1.48)	4.62 (1.71)	5.98 (1.10)
RUS	1.95 (1.10)	5.93 (1.15)	2.67 (1.67)	2.83 (1.68)	4.48 (1.61)	5.60 (1.28)
MEX	1.75 (1.09)	6.18 (0.95)	2.76 (1.53)	2.51 (1.52)	4.51 (1.65)	5.92 (1.10)
IND	1.71 (1.02)	6.17 (1.02)	2.54 (1.67)	2.33 (1.49)	4.56 (1.62)	5.92 (1.12)
IDN	1.99 (1.16)	6.12 (0.88)	3.06 (1.70)	2.65 (1.43)	4.62 (1.40)	5.86 (1.11)
PHL	1.53 (0.84)	6.09 (1.05)	2.71 (1.63)	2.00 (1.30)	4.63 (1.63)	5.92 (1.09)
VEN	1.31 (0.62)	6.27 (0.98)	2.76 (1.56)	2.14 (1.43)	4.97 (1.66)	5.97 (1.09)

This table shows the manifest scale means and standard deviations for all groups and factors.

Table 4. Manifest Correlations between Factors

	Amotivation	Material	Social	Introjected	Identified
Material	-0.21^{***}				
Social	0.02	0.12^{***}			
Introjected	0.13^{***}	0.08^{***}	0.52^{***}		
Identified	-0.18^{***}	0.35^{***}	0.43^{***}	0.39^{***}	
Intrinsic	-0.43^{***}	0.31^{***}	0.28^{***}	0.20^{***}	0.46^{***}

This table shows the Pearson correlations between the observed scores of the six factors of the MCMS, calculated on the total sample (N = 5857). $^{***}p < 0.001$.

both samples) and CFI was below 0.95.[25] In contrast, the MCMS achieved good fit measures on the total sample and for the individual countries (with CFIs above 0.95 for most countries and an RMSEA below 0.06 for all countries).

Compared to `A-WEIMS-M3`, the model that had the best fit in our evaluation of traditional work motivation scales, the MCMS showed a similar fit and suffers from none of `A-WEIMS-M3`'s drawbacks described in Section 3. Moreover, if only the four constructs measured by `A-WEIMS-M3` are taken into account, then the MCMS achieves a better fit: A four-factor model of the MCMS with the intrinsic motivation and social external regulation factors omitted had a CFI of 0.983, a TLI of 0.984, an RMSEA of 0.037 and an SRMR of 0.026 on the total sample (N = 5857) and a CFI of 0.979, a TLI of 0.971, an RMSEA of 0.039 and an SRMR of 0.035 on the USA sample (N = 721).

Construct Validity. Construct validity refers to the extent to which the items of a scale "accurately reflect the theoretical latent constructs they are designed to measure" [40]. In the context of confirmatory factor analysis, a poor model fit is considered evidence of a lack of construct validity. Besides achieving a good model fit, there are four additional components of construct validity that can be evaluated: content validity, convergent validity, discriminant validity, and nomological validity [40]. Content validity was ensured during item development (see Section 4). Evidence of

[25] TLI and SRMR were not reported.

Table 5. Internal Consistency of the MCMS

Group	Amotivation	Material	Social	Introjected	Identified	Intrinsic
ALL	0.78 (0.77 0.79)	0.78 (0.77 0.79)	0.84 (0.83 0.84)	0.83 (0.82 0.84)	0.87 (0.86 0.88)	0.88 (0.88 0.89)
HIGH	0.84 (0.82 0.85)	0.82 (0.80 0.83)	0.85 (0.84 0.86)	0.86 (0.85 0.87)	0.87 (0.86 0.88)	0.90 (0.89 0.91)
MID	0.70 (0.68 0.73)	0.75 (0.73 0.77)	0.82 (0.81 0.84)	0.83 (0.82 0.84)	0.86 (0.85 0.87)	0.87 (0.86 0.88)
LOW	0.78 (0.76 0.80)	0.77 (0.75 0.79)	0.85 (0.84 0.87)	0.80 (0.78 0.82)	0.85 (0.83 0.86)	0.87 (0.86 0.88)
USA	0.86 (0.84 0.87)	0.84 (0.82 0.86)	0.83 (0.81 0.85)	0.83 (0.81 0.85)	0.85 (0.84 0.87)	0.90 (0.89 0.91)
ESP	0.83 (0.81 0.86)	0.80 (0.78 0.83)	0.85 (0.83 0.87)	0.87 (0.86 0.89)	0.89 (0.88 0.91)	0.90 (0.89 0.92)
DEU	0.82 (0.80 0.85)	0.80 (0.78 0.83)	0.87 (0.86 0.89)	0.82 (0.79 0.85)	0.85 (0.82 0.87)	0.86 (0.84 0.88)
BRA	0.52 (0.45 0.59)	0.67 (0.62 0.72)	0.83 (0.80 0.85)	0.81 (0.78 0.84)	0.89 (0.87 0.90)	0.86 (0.84 0.88)
RUS	0.77 (0.74 0.80)	0.81 (0.78 0.83)	0.88 (0.87 0.90)	0.88 (0.86 0.90)	0.85 (0.83 0.87)	0.89 (0.88 0.90)
MEX	0.79 (0.76 0.82)	0.72 (0.68 0.75)	0.77 (0.74 0.80)	0.78 (0.75 0.81)	0.85 (0.83 0.87)	0.86 (0.84 0.88)
IND	0.74 (0.70 0.78)	0.82 (0.79 0.84)	0.87 (0.85 0.88)	0.79 (0.76 0.82)	0.85 (0.83 0.87)	0.87 (0.86 0.89)
IDN	0.79 (0.75 0.82)	0.65 (0.59 0.71)	0.85 (0.83 0.88)	0.79 (0.75 0.82)	0.79 (0.75 0.82)	0.84 (0.82 0.87)
PHL	0.80 (0.77 0.83)	0.80 (0.77 0.83)	0.84 (0.81 0.86)	0.82 (0.79 0.84)	0.88 (0.86 0.90)	0.88 (0.87 0.90)
VEN	0.60 (0.55 0.66)	0.77 (0.74 0.80)	0.76 (0.73 0.80)	0.77 (0.73 0.80)	0.86 (0.84 0.88)	0.83 (0.81 0.86)

This table shows Cronbach's alpha values for all groups and constructs, along with a 95% confidence interval for the values.

Table 6. Confirmatory Factor Analysis of the MCMS

Group	N	S-Bχ^2	df	CFI	TLI	RMSEA	RMSEA 90% CI	SRMR
ALL	5857	1590.49	120	0.965	0.955	0.046	0.044 0.048	0.037
HIGH	1952	573.64	120	0.970	0.961	0.044	0.041 0.047	0.037
MID	1834	557.07	120	0.964	0.955	0.045	0.041 0.048	0.038
LOW	1508	554.52	120	0.955	0.942	0.049	0.045 0.053	0.036
USA	721	281.45	120	0.965	0.956	0.043	0.037 0.049	0.040
ESP	677	272.66	120	0.975	0.968	0.043	0.037 0.050	0.040
DEU	554	284.42	120	0.955	0.943	0.050	0.043 0.056	0.044
BRA	496	256.58	120	0.957	0.946	0.048	0.040 0.055	0.044
RUS	677	311.67	120	0.966	0.957	0.049	0.043 0.055	0.039
MEX	661	316.27	120	0.948	0.933	0.050	0.043 0.056	0.048
IND	608	272.66	120	0.962	0.951	0.046	0.039 0.052	0.042
IDN	401	272.08	120	0.931	0.912	0.056	0.049 0.064	0.050
PHL	499	291.47	120	0.954	0.941	0.054	0.046 0.061	0.045
VEN	563	217.07	120	0.966	0.956	0.038	0.031 0.045	0.039

This table shows the goodness-of-fit statistics for the hypothesized MCMS model. The fit statistics are given for the total sample as well as for all groups.

Table 7. Estimated Loadings and Intercepts

	Amotivation		Material		Social		Introjected		Identified		Intrinsic	
	λ	τ	λ	τ	λ	τ	λ	τ	λ	τ	λ	τ
item 1	0.775	1.824	0.823	5.987	0.894	2.285	0.787	2.251	0.829	4.274	0.868	5.623
item 2	0.745	1.683	0.627	6.101	0.832	2.041	0.750	2.309	0.828	3.967	0.817	5.758
item 3	0.700	2.012	0.776	6.059	0.710	3.086	0.826	2.186	0.845	4.560	0.855	5.637

This table shows the standardized item loadings (λ) and the item intercepts (τ) estimated by the hypothesized model (N = 5857). The first item of each construct is the marker item, with its (unstandardized) loading fixed at 1. The order of the items follows the order in Table 14.

Table 8. Variance Extracted and Shared Variance

	Amotivation	**Material**	**Social**	**Introjected**	**Identified**	**Intrinsic**
AVE	0.548	0.557	0.665	0.621	0.696	0.718
MSV	0.274	0.194	0.360	0.360	0.276	0.276

This table shows the average variance extracted (AVE) and maximum shared variance (MSV) between the constructs. An AVE ≥ 0.5 indicates good convergence, and an AVE > MSV is evidence for discriminant validity.

the validity of the MCMS with respect to convergent, discriminant and nomological validity is discussed below. We follow the definitions of Hair et al. [40] for these types of validity.

Convergent Validity is established if the items measuring a construct "converge or share a high proportion of variance in common" [40]. All factor loadings should be statistically significant and, especially for large samples, should be at least ≥0.5 (ideally, ≥0.7) [40]. Table 7 shows the standardized factor loadings for each construct in the MCMS. All factor loadings, except for the item ExMat2 with a loading of 0.627, were 0.7 or higher and all loadings were statistically significant at $p < 0.001$, which provides evidence of convergent validity.

In addition to inspecting the factor loadings, the average variance extracted (AVE) can be used as a summary indicator of convergence. The AVE measures the average percentage of variation in the items explained by their common latent variable. AVE is calculated as the mean of the squared standardized factor loadings of a construct. An AVE of ≥0.5 is considered to indicate good convergence [30, 40]. Table 8 shows the AVE for each construct. For all constructs, the AVE was ≥0.5, providing further evidence of convergent validity.

Discriminant Validity is "the extent to which a construct or variable is truly distinct from other constructs or variables" [40]. To establish discriminant validity, each construct's AVE should be greater than its squared correlations with other constructs, indicating that more variance in the construct's items is explained by the construct than the construct shares with other constructs [30, 40]. The highest squared correlation of a construct with any other construct, or maximum shared variance (MSV), is shown in Table 8. For all constructs, the AVE was larger than the MSV, providing evidence of discriminant validity.

Nomological Validity is concerned with whether the relationships between the constructs of the scale correspond to theory and prior research [40]. Table 9 shows the estimated correlations between the constructs of the MCMS. As hypothesized by SDT and found in previous studies (e.g., References [35, 88]), we observe a negative correlation between intrinsic motivation and amotivation (as well as a moderate negative correlation between identified regulation and amotivation). Furthermore, as hypothesized by SDT, the strongest correlations are between adjacent constructs (intrinsic motivation with identified regulation and, introjected regulation with social external regulation), and generally, constructs tend to correlate more strongly with adjacent constructs

Table 9. Estimated Correlations between Constructs

	Amotivation	Material	Social	Introjected	Identified
Material	-0.263^{***}				
Social	0.051^{**}	0.128^{***}			
Introjected	0.151^{***}	0.108^{***}	0.600^{***}		
Identified	-0.222^{***}	0.440^{***}	0.458^{***}	0.449^{***}	
Intrinsic	-0.523^{***}	0.362^{***}	0.283^{***}	0.230^{***}	0.525^{***}

This table shows the model estimates of the correlations between constructs (N = 5857). $^{**}p < 0.01$, $^{***}p < 0.001$.

than with non-adjacent constructs. There are exceptions to this pattern; notably concerning the material external regulation construct, which has a stronger correlation with intrinsic motivation and identified regulation than with adjacent constructs. However, deviations from the pattern of ordered correlations are not uncommon in SDT-based motivation scales (e.g., References [11, 46, 58, 68]). Research suggests that if a continuum structure of motivation exists, then its nature is not necessarily in line with the assumptions of SDT or represented well by the pattern of correlations estimated by CFA (e.g., References [46, 58]). Therefore, further research is necessary to determine the extent to which the MCMS follows a continuum structure and to what extent this continuum structure is in line with the nature of the continuum hypothesized by SDT. Furthermore, while SDT hypothesizes that external regulation and intrinsic motivation are unrelated, previous research found a significant positive correlation between these constructs for both the MWMS (0.11) and the Academic Motivation Scale (0.49) [11]. In line with these previous empirical results, our results also show a positive correlation between intrinsic motivation and external regulation.

Criterion Validity. For a first analysis of the relationship between crowdworkers' motivations and their behavioral outcomes, we investigated the relationships between the MCMS's motivational constructs and (1) the time taken to complete the task, (2) the amount of text content that crowdworkers produced in response to the open-ended question, and (3) the self-reported time spent on CrowdFlower per week. We used the first two measures (*time taken* and *text produced*) as an estimate for the effort that workers put into the task.

Previous empirical research found that amotivation correlates negatively with effort [35], that autonomous motivation[26] correlates more positively with effort than controlled motivation[27] [35, 84], and that autonomous motivation predicts greater effort [84]. These previous findings are consistent with SDT: For intrinsic motivation, the positive emotions associated with enjoying an activity naturally reinforce persistent effort, and identified regulation may elicit such reinforcing emotions due to value congruence even if the activity itself is not enjoyable [84]. In line with previous research [35] and consistent with SDT, we expected effort to be correlated negatively with amotivation and positively with autonomous motivation. Furthermore, we expected the correlation of effort with autonomous motivation to be more positive than the correlation of effort with controlled motivation.

Concerning the weekly time spent on the platform (*weekly time*), in line with previous research on crowdworker motivations [51], we expected the motivational profiles to differ between workers who spend a lot of time on the platform and workers who spend little time on the platform.

[26] Autonomous motivation encompasses identified regulation and intrinsic motivation (e.g., Reference [35, 84]).
[27] Controlled motivation encompasses external regulation and introjected regulation (e.g., Reference [35, 84]).

Table 10. Correlations between Motivational Constructs and Estimates of Effort

	Amotivation	**Material**	**Social**	**Introjected**	**Identified**	**Intrinsic**
time taken	-0.14^{***}	0.04^{**}	0.09^{***}	0.10^{***}	0.18^{***}	0.15^{***}
text produced	-0.14^{***}	0.06^{***}	0.00	0.03^{*}	0.11^{***}	0.10^{***}
weekly time	-0.15^{***}	0.13^{***}	0.14^{***}	0.08^{***}	0.21^{***}	0.24^{***}

This figure shows the Pearson correlations between the different types of motivations and two estimates or effort: the time taken to complete the task and the amount of text content produced by the workers. Furthermore, it shows the Spearman correlations between the different types of motivation and weekly time spent on the platform. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Specifically, Kaufmann et al. [51] found that skill variety,[28] human capital advancement[29] and community identification[30] correlated positively with weekly time spent, while the correlation with pastime[31] was negative. Furthermore, in the context of SDT, previous research (e.g., Reference [23]) found that intrinsic motivation and highly internalized extrinsic motivation was associated with longer persistence in an activity. Therefore, we expected amotivation to be negatively correlated with weekly time spent, and we expected the strongest positive correlations to be with identified regulation and intrinsic motivation.

To calculate the time taken to complete the task, we used the starting and finishing times reported by the platform, and for an estimate of the amount of text content that a worker had produced, we counted the characters typed after removal of stopwords. Table 10 shows the Pearson correlations between the different types of motivation and the two estimates of effort.[32] While our estimates are noisy operationalizations of job effort (e.g., due to differences in reading and typing skills), they measure observed behavior as opposed to self-reports.

Our results support the hypothesis regarding effort, and we observe the same pattern as found in previous research [35] via self-reported job effort: Both measures have a significant negative correlation with amotivation and significant positive correlations with autonomous types of motivation (identified regulation and intrinsic motivation). Furthermore, the time taken correlates positively with controlled motivation types, and the amount of text produced correlates positively with material external regulation and introjected regulation. As hypothesized, the correlations between effort and autonomous motivation are more positive than the correlations between effort and controlled motivation. All differences in the strengths of the correlations between the types of autonomous motivation and the types of controlled motivation are statistically significant.[33]

[28] Kaufmann et al. [51] defined skill variety as "usage of a diversity of skills that are needed for solving a specific task and fit with the skill set of the worker," e.g., a worker picking "a translation task, because he likes translating." In the SDT context, this construct contains aspects of intrinsic motivation and the fulfillment of the need for competence.

[29] Kaufmann et al. [51] defined human capital advancement as "motivation through the possibility to train skills that could be useful to generate future material advantages," e.g., a worker choosing a task, "because he or she wants to improve language skills for a new or better job," which can be interpreted as an aspect of identified regulation.

[30] Kaufmann et al. [51] defined community identification as a "subconscious adoption of norms and values from the crowdsourcing platform community, which is caused by a personal identification process." In the SDT context, this construct is most similar to a regulation that has been internalized to a great extent, i.e., identified or integrated regulation.

[31] Kaufmann et al. [51] defined pastime as "acting just to kill time," e.g., a worker who "works on various 'random' tasks, because he has nothing better to do." In the context of SDT, this could be interpreted as amotivation.

[32] We logarithmized the character count as the distribution was heavily skewed.

[33] To assess the statistical significance of the differences, we used Steiger's method [86] for statistical comparisons between correlations measured on the same sample, as implemented by Lee and Preacher [56]. Our results showed that the correlations between the types of autonomous motivation and both measures of task effort were significantly stronger than the correlations between these measures and the types of controlled motivation. The differences were significant at $p < 0.01$ or lower.

To gather data on the weekly time spent on CrowdFlower, we asked workers the question *"How much time do you spend on CrowdFlower, per week?"* and workers were given seven answer possibilities, ranging from "less than 1 hour" to "more than 40 hours." This question was included in the demographics part of the task. Table 10 shows the correlations between the motivational constructs and the weekly time spent on CrowdFlower. Due to the ordinal nature of this variable, we report Spearman correlations. In line with previous research on SDT as well as previous research on crowdworker motivations, we find the highest positive correlations of weekly time spent with identified regulation and intrinsic motivation, and we find a negative correlation with amotivation.

Applicability Across Platforms. For a first evaluation of the extent to which the MCMS validly measures motivations of crowdworkers on other microtask platforms, we administered the MCMS to a small sample of crowdworkers on AMT. For use on AMT, we substituted the term "CrowdFlower tasks" with "tasks on Amazon Mechanical Turk" in the stem and items of the MCMS. We collected 150 responses from Indian workers on AMT in June 2017. After spam removal, the sample contained 109 responses. CFA results on this sample showed good fit ($CFI = 0.961$, $TLI = 0.951$, $RMSEA = 0.049$, $SRMR = 0.068$). Furthermore, measurement invariance tests (see Section 6 for more details on measurement invariance) revealed that the Indian AMT sample showed scalar invariance with our subsample of Indian crowdworkers on CrowdFlower. These results indicate that not only is the MCMS likely to be valid on other platforms, but that the MCMS also likely allows for valid comparisons of the motivational dimensions' group means across the platforms.

In sum, our results indicate that the MCMS offers reliable and valid measures of crowdworker motivations within the framework of SDT. Researchers wishing to measure the motivation of crowdworkers, for example along with other variables such as behavioral patterns of crowdworkers, can easily include the scale as a module in their task design. Instruction for use of the MCMS are given in Appendix D.

6 CROSS-GROUP COMPARABILITY OF MCMS RESULTS

When comparing the results of a measurement instrument across different groups, it is important to ensure that the instrument possesses the same psychometric properties in all groups. This characteristic is referred to as *measurement invariance*. Tests of measurement invariance evaluate "whether or not, under different conditions of observing and studying phenomena, measurement operations yield measures of the same attribute" [43]. In our case, measurement invariance means that crowdworkers from different countries (or country income groups) assign the same meaning to the items used in the MCMS. Measurement invariance is particularly important if mean level differences between countries ought to be compared. A lack of measurement invariance can indicate, for example, that respondents of different groups understand the items in a different way (e.g., due to culture) or that different levels of response biases are present (e.g., Reference [13]).

To evaluate measurement invariance of the MCMS, we conducted multiple-group confirmatory factor analyses (MGCFA). In MGCFA, measurement invariance is evaluated via a series of hypothesis tests, which test invariance at different levels. Three levels of measurement invariance are commonly tested: configural, metric, and scalar invariance (e.g., References [12, 14]). Configural invariance requires that the items share the same configurations of loadings in all groups. Metric invariance additionally requires that the loadings of each item on its factor is the same across groups. Scalar invariance additionally requires that the intercepts of item regressions on each factor are the same across groups. To validly compare manifest mean differences across groups, scalar invariance is required. To validly compare latent mean differences across groups, at least partial scalar invariance is required [63].

We tested configural, metric and scalar invariance of the MCMS across countries and across income groups. Configural invariance is indicated by acceptable goodness-of-fit indices in an

Table 11. Measurement Invariance

	CFI	CFI Δ	RMSEA	RMSEA Δ
Income Groups				
Configural Invariance	0.964	n/a	0.046	n/a
Metric Invariance	0.963	0.001	0.045	0.001
Full Scalar Invariance	0.952	0.011	0.049	0.005
Partial Scalar Invariance*	0.955	0.008	0.048	0.004
Countries				
Configural Invariance	0.960	n/a	0.047	n/a
Metric Invariance	0.959	0.001	0.046	0.001
Full Scalar Invariance	0.930	0.028	0.058	0.011
Partial Scalar Invariance**	0.952	0.007	0.049	0.002

This table shows the results of the MGCFA tests of invariance between groups and countries. The deltas are with respect to the previous level of measurement invariance, i.e., for scalar invariance, the sum of the deltas for metric and scalar invariance should be below 0.01 for CFI and below 0.015 for RMSEA. *One free intercept (*Am3*); **five free intercepts (*Am3*, *ExMat2*, *ExSoc3*, *Introj2*, and *Ident2*).

MGCFA model without any equality constraints [91]. For indications of metric and scalar non-invariance, we follow the guidelines of Chen [12]: For metric and scalar invariance tests on large samples ($N > 300$), a change of ≥ -0.010 in CFI supplemented by a change of $\geq$0.015 in RMSEA indicates non-invariance.

Table 11 shows the goodness-of-fit indices for the progressively restricted models. The results show a good fit for the model without equality constraints, indicating that configural invariance holds. Full metric invariance was also achieved, indicating that the strength of the relationship between the items and constructs is the same across groups. Full scalar invariance could not be achieved. However, partial scalar invariance was achieved by releasing one intercept for the income groups (*Am3*) and five intercepts in the countries (*Am3*, *ExMat2*, *ExSoc3*, *Introj2*, and *Ident2*). Partial scalar invariance still allows for factor means to be compared as long as at least two intercepts per factor are invariant. Care should be taken, however, as Steinmetz' [87] simulations showed that unequal intercepts may lead to erroneous conclusions about mean-level differences when comparing observed composite means across groups. Therefore, a cross-country or cross-income group comparison of crowdworker motivations measured with the MCMS should rely on the model-implied latent means, which take intercept non-invariance into account, instead of observed composite means.

In sum, our analyses showed that the MCMS is well-suited for measuring the motivations of crowdworkers based in different countries. Furthermore, the invariance analyses indicated that the motivations measured with the MCMS can be used for a comparison of motivations across countries and across country income groups.

7 MOTIVATIONS OF CROWDWORKERS ON CROWDFLOWER

This section reports the motivations of crowdworkers on CrowdFlower as measured by the MCMS as well as the results of our cross-country and cross-income group comparison of crowdworker motivations.

Motivations Measured with the MCMS. Table 13 shows the latent means[34] of the different motivational dimensions for the entire sample (ALL). Our results show that overall, material

[34] To obtain estimates of the latent means, we fixed the marker items' intercepts to zero.

Table 12. Responses to the Open-ended Question

Construct	Examples
Amotivation	*"I'm bored," "I have nothing else to do," "this is boring," "Nothing."*
Material	*"I need the money," "I get paid," "To get an extra income!," "good profit."*
Social	*"other people want me to fulfill the job," "my friends do ti too," "referral from a friend," "friends do it."*
Introjected	*"I'm no worse than others," "To feel better about myself," "I feel useful," "They make me feel productive and useful."*
Identified	*"To be my own boss!," "To gain some new experience," "It helps me improve my English," "I can work from anywhere."*
Intrinsic	*"Enjoy seeing a variety of different topics," "most of the tasks are actually entertaining," "There are taske that are very interesting," "I enjoy sharing thoughts and opinions while completing CrowdFlower tasks."*

This table shows illustrative examples of answers to the open-ended question "Give five reasons for why you do CrowdFlower tasks" from crowdworkers who had a mean score of >5 on the respective construct.

external regulation was the most important motivation for crowdworkers, with a mean of 5.99, followed by intrinsic motivation with a mean of 5.62. This points to an interesting duality of monetary and interest-driven, enjoyment-based motivational influences. The result is consistent with previous research on crowdworker motivations on AMT (e.g., Reference [51]), which found that both payment and enjoyment play an important role for crowdworkers, with monetary reasons being slightly more important than enjoyment. The construct with the third highest mean was identified regulation (mean 4.27), which signifies that putting effort into CrowdFlower tasks is moderately in alignment with crowdworkers' personal goals and objectives such as lifestyle preferences or career plans. Social external regulation, introjected regulation and amotivation were the least important motivational factors for crowdworkers overall, with amotivation having the lowest score of all motivational dimensions (means 2.29, 2.25, and 1.82, respectively).

As an illustration of how the different motivational types can be interpreted in the microtask context, we give examples of the reasons that crowdworkers have for doing CrowdFlower tasks in the workers' own words. In the first section of the task, we instructed crowdworkers to think of five reasons for why they do tasks on CrowdFlower and asked them to write down these reasons. For each construct, Table 12 shows examples of answers to this question that can be interpreted to correspond to the theoretical definition of the construct. The examples given are taken, for each construct, from four different crowdworkers who had a mean score greater than 5 on the respective construct.

Differences in Motivations across Groups. As partial scalar invariance was achieved, the analysis of group differences in motivations measured with the MCMS relies on latent means estimated by the model instead of observed means. For analyzing the differences in latent constructs between groups, one group was chosen as the reference group. For our analysis, we chose the high-income group sample as the reference group for the cross-income group comparison and the USA sample as the reference group for the cross-country comparison. Figure 3 shows the differences in latent means of the different countries, compared to the reference group (USA), and Table 13 shows the latent means for all groups, as well as the mean differences to the reference group in parentheses.[35]

[35] We obtained the estimated differences and their statistical significances via MGCFA with the means of the reference group fixed to zero. The threshold for all reported significances is set at $p < 0.001$, except amotivation in Spain, external social regulation in Germany and external material regulation in Russia ($p < 0.05$).

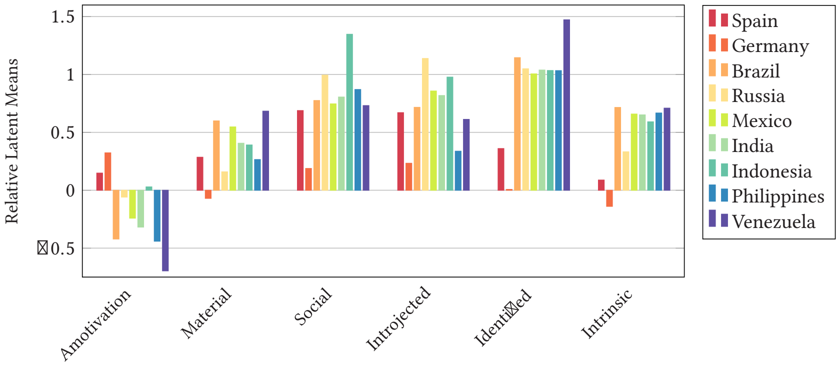


Fig. 3. **Country differences in latent means, compared to the USA sample.** This figure shows the differences in latent means for all constructs and countries, compared to the latent means of the U.S. sample.

Table 13. Latent Scale Means and Group Differences

Group	Amotivation		Material		Social		Introjected		Identified		Intrinsic	
ALL	1.82		5.99		2.29		2.25		4.27		5.62	
HIGH	2.12		5.75		1.86		1.92		3.61		5.22	
MID	1.75	(−0.36)	6.08	(0.32)	2.50	(0.64)	2.57	(0.65)	4.54	(0.93)	5.78	(0.56)
LOW	1.71	(−0.41)	6.06	(0.30)	2.58	(0.73)	2.30	(0.38)	4.60	(0.99)	5.87	(0.65)
USA	1.97		5.68		1.60		1.62		3.49		5.23	
ESP	2.12	(0.15)	5.97	(0.28)	2.29	(0.69)	2.29	(0.67)	3.85	(0.36)	5.31	(0.09)
DEU	2.30	(0.32)	5.61	(−0.07)	1.79	(0.19)	1.85	(0.23)	3.50	(0.01)	5.09	(−0.14)
BRA	1.55	(−0.42)	6.28	(0.60)	2.38	(0.78)	2.34	(0.72)	4.64	(1.15)	5.94	(0.72)
RUS	1.92	(−0.06)	5.84	(0.16)	2.60	(0.99)	2.76	(1.14)	4.54	(1.05)	5.56	(0.33)
MEX	1.73	(−0.24)	6.23	(0.55)	2.35	(0.75)	2.48	(0.86)	4.50	(1.00)	5.88	(0.66)
IND	1.66	(−0.32)	6.09	(0.41)	2.41	(0.80)	2.44	(0.82)	4.53	(1.04)	5.88	(0.65)
IDN	2.00	(0.03)	6.07	(0.39)	2.95	(1.35)	2.60	(0.98)	4.53	(1.03)	5.82	(0.59)
PHL	1.53	(−0.44)	5.95	(0.26)	2.47	(0.87)	1.96	(0.34)	4.53	(1.03)	5.89	(0.67)
VEN	1.28	(−0.70)	6.36	(0.68)	2.33	(0.73)	2.23	(0.61)	4.96	(1.47)	5.93	(0.71)

This table shows the latent means for all groups, as well as the estimated differences in latent means in parentheses.

Regarding the ranks of the constructs, the results show that the ranks of the three constructs with the highest scores were the same across all groups. External material regulation was the construct with the highest score, followed by intrinsic motivation with only a slightly lower score, in all countries and income groups. Furthermore, the third most important motivational factor was identified regulation in all groups. The other motivational dimensions differed in rank across countries and income groups. Amotivation was the construct with the lowest score in the middle- and low-income groups, while in the high-income group, it had a higher score than social external regulation and introjected regulation.

Regarding the differences in construct scores, the results show that motivations differ significantly between crowdworkers of different countries and country income groups.[36] The largest

[36] Note that, as described in Section 5, some care should be taken when interpreting results from Mexico and Indonesia due to a CFI lower than 0.95 (but above 0.90) and when interpreting the amotivation construct in Brazil and Venezuela, as well as the material external regulation construct in Brazil and Indonesia due to Cronbach's alpha values below 0.7.

differences in motivation, compared to workers in the USA, are with countries that are in income groups lower than the USA.

Amotivation was significantly higher in the high-income group than in the middle and low-income groups. Crowdworkers in Brazil, Mexico, India, the Philippines, and Venezuela had a significantly lower amotivation score than U.S. workers, while German and Spanish crowdworkers exhibited a significantly higher level of amotivation than workers in the USA. This indicates that in high income countries, crowdworkers tend to perceive doing CrowdFlower tasks as more "pointless" and a "waste of time" than in most lower income countries.

While material external regulation had the highest score of all constructs in all countries, scores were significantly higher in some countries than in others. Germany was the country with the lowest score on this construct, while Venezuela had the highest score. Both the middle and the low-income group reported higher scores for material external regulation than the high-income group, and crowdworkers of all countries except Germany reported a significantly higher material external regulation than U.S. workers. This means that workers in countries with lower incomes tend to be more motivated by the material rewards of microtasks than workers in high income countries.

Social external regulation and introjected regulation scores were significantly higher in the middle and low-income groups than in the high-income group, and significantly higher in all countries, compared to the USA sample. This means that satisfying external social demands, as well as the avoidance of shame or guilt feelings, are more important motivational factors for workers in low and middle income countries and in countries other than the USA.

The largest differences in construct scores, both between countries and between income groups, were found for identified regulation. Scores were significantly higher in countries of the middle and low-income groups than in countries of the high-income group. This means that in lower income countries, crowdworkers perceive putting effort into microtasks as more in line with their personal goals, objectives and values. One reason for this might be that in higher income countries, the same goals can be achieved more effectively through other means. Crowdworkers in the USA had the lowest identified regulation score of all countries and scores were significantly higher in all other countries except Germany.

Finally, all countries except Spain and Germany reported a significantly higher intrinsic motivation than workers in the USA, and middle and low-income groups reported a higher intrinsic motivation score than the high-income group. This means that that workers of countries in the middle and low-income groups are more driven by interest and enjoyment inherent in the activity.

8 CONCLUSION

In this article, we developed and validated the Multidimensional Crowdworker Motivation Scale (MCMS), a new instrument for measuring crowdworker motivations. The MCMS measures the motivations of crowdworkers on six dimensions, based on the conceptualization of motivation suggested by self-determination theory. To the best of our knowledge, the MCMS is the first instrument developed specifically for measuring motivation in the microtask context that provides a comprehensive representation of the motivational dimensions hypothesized by self-determination theory. Compared to existing instruments, the MCMS allows for a more comprehensive and theoretically well-founded measurement of crowdworker motivations with only three items per motivational dimension. Moreover, the MCMS is the first instrument for measuring crowdworker motivations that is validated in multiple countries and income groups.

The hypothesized six-dimensional model of the MCMS generally showed good fit in all countries and income groups. In addition, the results of the measurement invariance tests demonstrated that partial scalar invariance holds. This implies that the substantive meaning of the motivations measured with the MCMS are comparable across countries and income groups, allowing for valid cross-national comparisons. By providing a reliable scale for measuring crowdworker motivations, our study constitutes an important step towards a better understanding of the international crowdworkforce.

We designed the MCMS with generalizability across platforms in mind. Our initial results on a sample of crowdworkers on AMT support the scale's applicability to other microtask platforms beyond CrowdFlower. By exchanging the platform name in the instructions and items of the MCMS, the scale can be deployed to measure the motivations of crowdworkers on other microtask platforms as well. Moreover, it likely allows for valid cross-platform comparisons of motivations, as we demonstrated with a sample of workers on AMT.

Finally, in this article, we have presented a first cross-country and income group comparison of crowdworker motivations on the microtask platform CrowdFlower. This data provides novel insights regarding wide-ranging differences of the motivations for participating on such a platform.

Given the good psychometric properties of the MCMS, future research on crowdwork can utilize the scale to address substantive questions concerning crowd employment. Here, the six motivational dimensions could serve as an outcome or as an explanatory variable. Potential fields of application[37] include predicting worker retention, investigating the relationship between worker motivation and productivity in different tasks, and conducting comparison studies of the motivations of different crowdworker populations, among others. Furthermore, the MCMS contributes to answering the question as to where in the employment space crowd employment should be located. The MCMS is also relevant for microtask platform developers who can use the scale to assess whether changes made to the platform lead to desirable or undesirable changes in motivation.

The work presented in this article has several limitations. First, the scale was presented in English to the crowdworkers in all countries, which means that we are only able to capture the motivations of crowdworkers with appropriate English skills. However, we can assume that a majority of crowdworkers on CrowdFlower possess an adequate level of English skills, as the platform interface is available exclusively in English and workers are expected to understand instructions in English. Furthermore, demand for crowdworkers is driven by English-speaking countries [55]. Regarding the presence of social desirability bias, Blais et al. [6] found that self-reported work motivations only correlated very weakly with the Marlow-Crowne Social Desirability Scale [18]. However, as Antin and Shaw [1] found evidence for the presence of social desirability bias in self-reported motivations of crowdworkers, further experiments are needed to assess the extent to which social desirability bias is present in data collected with the MCMS. A further limitation of the MCMS is that it does not measure integrated regulation. The lack of an integrated regulation factor in the MCMS is due to problems of statistically distinguishing this factor from identified regulation and intrinsic motivation. Due to the same problems, this limitation also applies to other SDT-based work motivation scales such as the MWMS [35]. Finally, the MCMS was developed specifically for the context of paid microtasks and is not intended for use in other crowdsourcing contexts. While an application in other contexts may be possible, the scale would have to be adapted first (e.g., by removing the material external regulation construct for application in the context of unpaid tasks) and validated in the respective context.

In future work, we plan to conduct a more in-depth analysis of cross-country and cross-income group differences, including their stability over time. Additionally, we plan to further investigate the relationship between motivations and economic as well as demographic factors, going beyond the country of residence as an indicator of difference. Regarding cross-platform comparability of MCMS responses, we plan to further evaluate the MCMS on other microtask platforms and,

[37] For instructions on use of the MCMS, see Appendix D.

provided that measurement invariance is achieved, conduct a cross-platform analysis of worker motivations.

Another direction that we plan to follow in future work is an evaluation of the extent to which motivations measured with the MCMS are related to different antecedents and outcomes. For investigating the relations between measured motivations and antecedents (e.g., the satisfaction of basic needs) or outcomes (e.g., emotional exhaustion), the scales for measuring the antecedents and outcomes will first have to be validated within the crowdworking domain. Finally, in future work, we plan to develop Bayesian models that incorporate not only the responses to the MCMS items but also responses to open-ended survey questions and demographic metadata.

To conclude, the MCMS constitutes a promising step forward in measuring the motivations of crowdworkers in a theoretically founded, reliable and internationally comparable way. By shedding light on the motivations of the "indefinite and unknown group" of crowdworkers, the MCMS enables novel insights into this emerging form of short-term employment. This work is relevant not only for researchers but also for practitioners seeking to measure the motivations of crowdworkers and to harness knowledge about differences in crowdworker motivations. Further information about our work is available on our website [29].

APPENDIX

A THE MULTIDIMENSIONAL CROWDWORKER MOTIVATION SCALE

Table 14. The Multidimensional Crowdworker Motivation Scale

		Source
Amotivation		
Am1	*I don't know why, CrowdFlower tasks often seem like a waste of time.*	
Am2	*I don't know why I'm doing CrowdFlower tasks, it's pointless work.*	[35]
Am3	*I don't know why, I often perceive CrowdFlower tasks as an annoying chore.*	[26]
External Regulation (Material)		
ExMat1	*Because CrowdFlower tasks give me financial gains.*	
ExMat2	*For the income CrowdFlower tasks provide me.*	[88]
ExMat3	*Because of the money I get from doing CrowdFlower tasks.*	
External Regulation (Social)		
ExSoc1	*Because other people want me to do CrowdFlower tasks (e.g., family, friends,...).*	
ExSoc2	*Because other people say I should (e.g., family, friends,...).*	[67]
ExSoc3	*Because other people expect it of me (e.g., family, friends,...).*	
Introjected Regulation		
Introj1	*Because otherwise I would have a bad conscience.*	[26]
Introj2	*Because otherwise I will feel ashamed of myself.*	[35]
Introj3	*Because otherwise I will feel bad about myself.*	[35]
Identified Regulation		
Ident1	*Because this is the type of work I chose to do to attain a certain lifestyle.*	[88]
Ident2	*Because I chose this type of work to attain my career goals.*	[88]
Ident3	*Because it is the type of work I have chosen to attain certain important objectives.*	[88]
Intrinsic Motivation		
Intrin1	*Because I have fun doing CrowdFlower tasks.*	[35]
Intrin2	*Because I enjoy doing CrowdFlower tasks.*	[67]
Intrin3	*Because what I do in CrowdFlower tasks is interesting.*	[35]

The stem is "Why do you or would you put efforts into doing CrowdFlower tasks?" (adapted from MWMS [35]). All items were answered on a 7-point Likert scale ranging from "not at all" (1) to "completely" (7). The column "Source" indicates from which motivation scale the item was adapted.

B MCDONALD'S COEFFICIENT OMEGA

As an alternative to Cronbach's alpha, we additionally calculated McDonald's coefficient omega [62] for assessing the reliability of the MCMS. Table 15 displays the values of omega for each country and income group. As with Cronbach's alpha, in most countries and groups, omega exceeds 0.7 for each construct. Exceptions to this are the amotivation factor in Brazil and Venezuela as well as the material external regulation factor in Brazil and Indonesia with values between 0.5 and 0.7. Compared to Cronbach's alpha, the values of coefficient omega are equal or slightly higher.[38]

Table 15. McDonald's Coefficient Omega

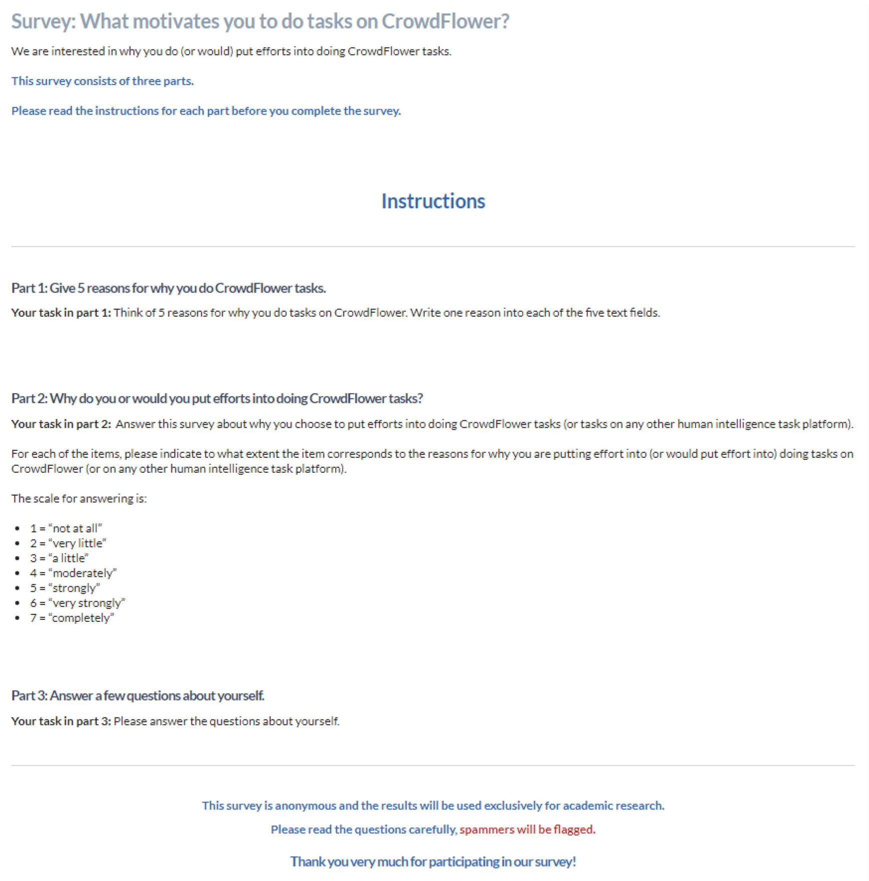
Group	Amotivation	Material	Social	Introjected	Identified	Intrinsic
ALL	0.78 (0.77 0.80)	0.79 (0.78 0.81)	0.84 (0.83 0.85)	0.83 (0.82 0.84)	0.87 (0.86 0.88)	0.88 (0.88 0.89)
HIGH	0.84 (0.82 0.86)	0.82 (0.80 0.84)	0.86 (0.84 0.88)	0.86 (0.84 0.88)	0.87 (0.85 0.88)	0.89 (0.88 0.91)
MID	0.71 (0.67 0.74)	0.76 (0.74 0.79)	0.83 (0.81 0.84)	0.83 (0.81 0.85)	0.86 (0.85 0.88)	0.88 (0.86 0.89)
LOW	0.78 (0.74 0.81)	0.78 (0.74 0.81)	0.86 (0.84 0.87)	0.80 (0.77 0.83)	0.85 (0.83 0.87)	0.87 (0.86 0.89)
USA	0.86 (0.83 0.88)	0.84 (0.81 0.88)	0.84 (0.80 0.88)	0.83 (0.79 0.88)	0.85 (0.83 0.88)	0.91 (0.89 0.92)
ESP	0.84 (0.81 0.87)	0.81 (0.78 0.85)	0.86 (0.83 0.88)	0.87 (0.85 0.90)	0.89 (0.88 0.91)	0.90 (0.89 0.92)
DEU	0.83 (0.8 0.86)	0.81 (0.78 0.85)	0.88 (0.85 0.91)	0.82 (0.78 0.87)	0.85 (0.82 0.87)	0.86 (0.84 0.89)
BRA	0.52 (0.44 0.61)	0.69 (0.62 0.76)	0.83 (0.80 0.86)	0.81 (0.77 0.85)	0.89 (0.87 0.91)	0.86 (0.83 0.90)
RUS	0.77 (0.73 0.81)	0.81 (0.78 0.85)	0.89 (0.87 0.91)	0.88 (0.86 0.90)	0.85 (0.83 0.88)	0.89 (0.87 0.91)
MEX	0.79 (0.74 0.84)	0.73 (0.68 0.79)	0.78 (0.74 0.81)	0.78 (0.74 0.82)	0.85 (0.83 0.87)	0.86 (0.83 0.89)
IND	0.74 (0.68 0.80)	0.82 (0.78 0.86)	0.87 (0.84 0.89)	0.79 (0.75 0.83)	0.85 (0.83 0.88)	0.88 (0.85 0.90)
IDN	0.79 (0.74 0.85)	0.65 (0.57 0.72)	0.86 (0.83 0.89)	0.79 (0.74 0.84)	0.79 (0.74 0.84)	0.85 (0.81 0.89)
PHL	0.81 (0.75 0.86)	0.80 (0.75 0.85)	0.84 (0.81 0.87)	0.82 (0.77 0.86)	0.88 (0.86 0.91)	0.89 (0.86 0.91)
VEN	0.63 (0.52 0.74)	0.77 (0.72 0.83)	0.77 (0.74 0.81)	0.77 (0.73 0.82)	0.86 (0.84 0.89)	0.84 (0.80 0.87)

This table shows McDonald's omega values for all groups and constructs, along with a 95% confidence interval for the values.

C TASK INTERFACE

The CrowdFlower task included a section in which workers were instructed to write down five reasons for why they do tasks on CrowdFlower, a section with the MCMS, and a section with questions about demographics and money use.

[38] One exception to this was intrinsic motivation in the high-income group, where coefficient omega is 0.01 lower than alpha.



Survey: What motivates you to do tasks on CrowdFlower?

We are interested in why you do (or would) put efforts into doing CrowdFlower tasks.

This survey consists of three parts.

Please read the instructions for each part before you complete the survey.

Instructions

Part 1: Give 5 reasons for why you do CrowdFlower tasks.

Your task in part 1: Think of 5 reasons for why you do tasks on CrowdFlower. Write one reason into each of the five text fields.

Part 2: Why do you or would you put efforts into doing CrowdFlower tasks?

Your task in part 2: Answer this survey about why you choose to put efforts into doing CrowdFlower tasks (or tasks on any other human intelligence task platform).

For each of the items, please indicate to what extent the item corresponds to the reasons for why you are putting effort into (or would put effort into) doing tasks on CrowdFlower (or on any other human intelligence task platform).

The scale for answering is:

- 1 = "not at all"
- 2 = "very little"
- 3 = "a little"
- 4 = "moderately"
- 5 = "strongly"
- 6 = "very strongly"
- 7 = "completely"

Part 3: Answer a few questions about yourself.

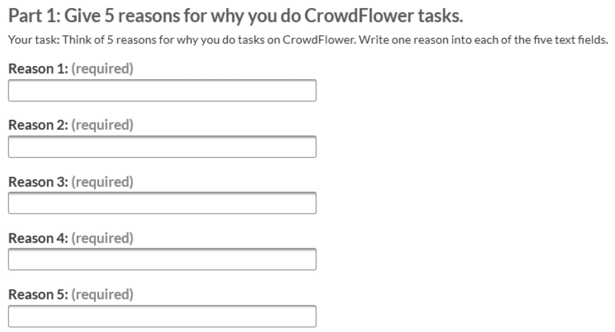
Your task in part 3: Please answer the questions about yourself.

This survey is anonymous and the results will be used exclusively for academic research.

Please read the questions carefully, spammers will be flagged.

Thank you very much for participating in our survey!

Fig. 4. **Task instructions.** This figure shows the task instructions that were shown to crowdworkers at the beginning of the task.



Part 1: Give 5 reasons for why you do CrowdFlower tasks.

Your task: Think of 5 reasons for why you do tasks on CrowdFlower. Write one reason into each of the five text fields.

Reason 1: (required)

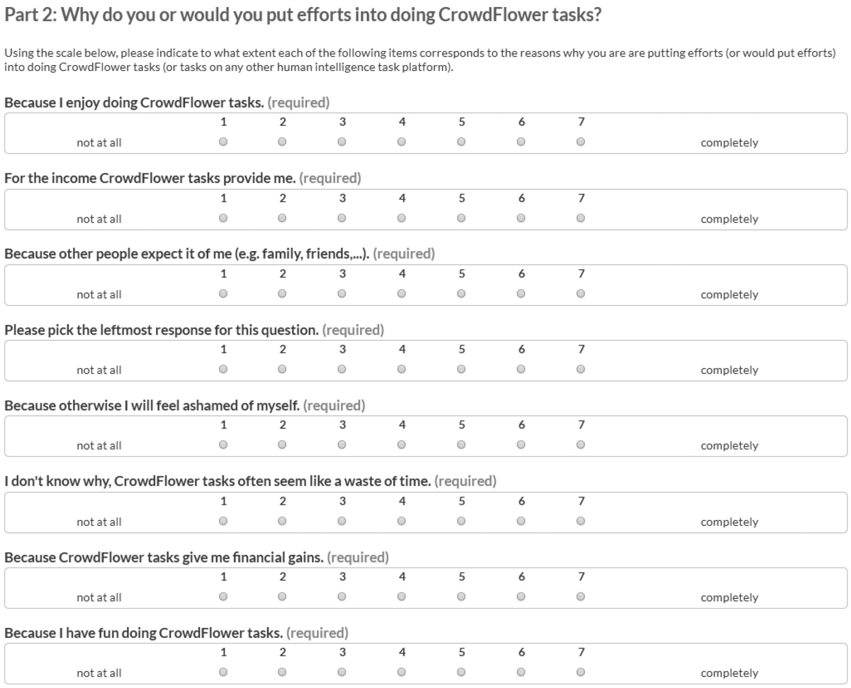
Reason 2: (required)

Reason 3: (required)

Reason 4: (required)

Reason 5: (required)

Fig. 5. **Interface for open-ended answers.** This figure shows the interface that crowdworkers were given to state five reasons for why they do tasks on CrowdFlower in their own words.



Part 2: Why do you or would you put efforts into doing CrowdFlower tasks?

Using the scale below, please indicate to what extent each of the following items corresponds to the reasons why you are are putting efforts (or would put efforts) into doing CrowdFlower tasks (or tasks on any other human intelligence task platform).

Because I enjoy doing CrowdFlower tasks. (required)

	1	2	3	4	5	6	7	
not at all	○	○	○	○	○	○	○	completely

For the income CrowdFlower tasks provide me. (required)

	1	2	3	4	5	6	7	
not at all	○	○	○	○	○	○	○	completely

Because other people expect it of me (e.g. family, friends,...). (required)

	1	2	3	4	5	6	7	
not at all	○	○	○	○	○	○	○	completely

Please pick the leftmost response for this question. (required)

	1	2	3	4	5	6	7	
not at all	○	○	○	○	○	○	○	completely

Because otherwise I will feel ashamed of myself. (required)

	1	2	3	4	5	6	7	
not at all	○	○	○	○	○	○	○	completely

I don't know why, CrowdFlower tasks often seem like a waste of time. (required)

	1	2	3	4	5	6	7	
not at all	○	○	○	○	○	○	○	completely

Because CrowdFlower tasks give me financial gains. (required)

	1	2	3	4	5	6	7	
not at all	○	○	○	○	○	○	○	completely

Because I have fun doing CrowdFlower tasks. (required)

	1	2	3	4	5	6	7	
not at all	○	○	○	○	○	○	○	completely

Fig. 6. **MCMS Interface.** This figure shows the interface of the task section in which crowdworkers answered the MCMS question by indicating agreement on the 18 items. Also note the included test question to check for spamming behavior and lack of attention. Due to space limitations, this screenshot shows only part of the scale. The full scale is shown in Appendix A.

D INSTRUCTIONS FOR USE OF THE MCMS

To measure the motivations of crowdworkers, researchers can administer the scale (shown in Appendix A) to their sample of workers. For example, the MCMS can be easily included as a module in the design of a microtask if a researcher wishes to measure motivation along with other variables, such as other characteristics or behavioral patterns, of their worker sample.

As long as the researcher does not wish to compare different groups of crowdworkers, the MCMS can be used as a summated scale. This means that the scores for the constructs can be obtained by averaging the scores of the individual items corresponding to each construct. For example, to obtain the score for the amotivation construct of a specific worker, the researcher takes the individual item responses of this worker for the items associated with the amotivation construct (i.e., Am1, Am2, and Am3) and calculates $Amotivation = (Am1 + Am2 + Am3)/3.0$.

An alternative way to obtain the construct scores is to specify the model shown in Figure 2 as a latent variable model, for example by using latent variable modeling software such as Mplus[39] or the R library lavaan.[40] This method[41] of obtaining the construct scores is preferable to averaging manifest item scores as it accounts for the measurement error necessarily present in the

[39] https://www.statmodel.com/.

[40] https://cran.r-project.org/web/packages/lavaan/lavaan.pdf.

[41] For an introduction to latent variable models and multivariate data analysis, we refer the reader to Kline [53], Hair et al. [40], and Bollen [7].

measurement of any abstract concept [40]. Furthermore, specifying the model allows researchers to conduct confirmatory factor analysis to validate the MCMS on their sample of crowdworkers, ensuring that the measurement is valid for their specific target population.

If a researcher wishes to compare the construct means between two groups of crowdworkers, such as male and female workers, workers of different age, workers on different platforms, or, as in the case of the present study, different countries, then the researcher has to ensure that the workers of the different groups assign the same meaning to the items used in the MCMS. This can be done by establishing measurement invariance between the groups as described in Section 6. Measurement invariance should be established before comparing the group means of any scale, especially when administering the scale to respondents of different cultures, to ensure that the differences in scores are not due to a different understanding of the items (e.g., due to culture) or due to measurement artifacts such as different levels of extreme or acquiescent response bias (e.g., Reference [13]).

REFERENCES

[1] Judd Antin and Aaron D. Shaw. 2012. Social desirability bias and self-reports of motivation: A study of Amazon mechanical turk in the U.S. and India. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'12)*, Joseph A. Konstan, Ed H. Chi, and Kristina Höök (Eds.). ACM, 2925–2934. DOI:https://doi.org/10.1145/2207676.2208699

[2] Peter M. Bentler and Douglas G. Bonett. 1980. Significance tests and goodness of fit in the analysis of covariance structures. *Psychol. Bull.* 88, 3 (1980), 588.

[3] Janine Berg. 2016. Income security in the on-demand economy: Findings and policy lessons from a survey of crowdworkers. *Comp. Lab. L. & Pol'y J.* 37 (2015), 543.

[4] Janine Berg, M. Furrer, E. Harmon, U. Rani, and M. S. Silberman. 2018. Digital labour platforms and the future of work: Towards decent work in the online world. *Geneva: International Labour Organization, September* 20 (2018).

[5] Michael S. Bernstein, Greg Little, Robert C. Miller, Björn Hartmann, Mark S. Ackerman, David R. Karger, David Crowell, and Katrina Panovich. 2015. Soylent: A word processor with a crowd inside. *Commun. ACM* 58, 8 (2015), 85–94. DOI:https://doi.org/10.1145/2791285

[6] M. R. Blais, L. Lachance, R. J. Vallerand, N. M. Briere, and A. S. Riddle. 1993. The work motivation inventory. *Revue Quebecoise de Psychologie* 14 (1993), 185–215.

[7] Kenneth A. Bollen. 1989. *Structural Equations with Latent Variables.* John Wiley & Sons.

[8] Robin Bordoli. 2018. Focused on the Future with a New Name. Retrieved from http://www.figure-eight.com/focused-future-new-name/.

[9] Dana Chandler and Adam Kapelner. 2013. Breaking monotony with meaning: Motivation in crowdsourcing markets. *J. Econ. Behav. Organiz.* 90 (2013), 123–133.

[10] Jonathan Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta (Eds.). Curran Associates, 288–296. Retrieved from http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.

[11] Emanuela Chemolli and Marylène Gagné. 2014. Evidence against the continuum structure underlying motivation measures derived from self-determination theory. *Psychol. Assess.* 26, 2 (2014), 575.

[12] Fang Fang Chen. 2007. Sensitivity of goodness of fit indexes to lack of measurement invariance. *Struct. Equat. Model.* 14, 3 (2007), 464–504.

[13] Gordon W. Cheung and Roger B. Rensvold. 2000. Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling. *J. Cross-Cultur. Psychol.* 31, 2 (2000), 187–212.

[14] Gordon W. Cheung and Roger B. Rensvold. 2002. Evaluating goodness-of-fit indexes for testing measurement invariance. *Struct. Equat. Model.* 9, 2 (2002), 233–255.

[15] Cristiano Codagnone, Fabienne Abadie, and Federico Biagi. 2016. The future of work in the "sharing economy." Market efficiency and equitable opportunities or unfair precarisation? *Market Efficiency and Equitable Opportunities or Unfair Precarisation* (2016).

[16] European Commission. 2016. A European agenda for the collaborative economy. Retrieved from http://ec.europa.eu/DocsRoom/documents/16881.

[17] Lee J. Cronbach. 1951. Coefficient alpha and the internal structure of tests. *psychometrika* 16, 3 (1951), 297–334.

[18] Douglas P. Crowne and David Marlowe. 1960. A new scale of social desirability independent of psychopathology. *J. Consult. Psychol.* 24, 4 (1960), 349.

[19] María de los Ángeles Morata-Ramírez and Francisco Pablo Holgado-Tello. 2013. Construct validity of Likert scales through confirmatory factor analysis: A simulation study comparing different methods of estimation based on Pearson and polychoric correlations. *Int. J. Soc. Sci. Studies* 1, 1 (2013), p54–61.

[20] Edward L. Deci and Richard M. Ryan. 1980. The empirical exploration of intrinsic motivational processes. *Adv. Exper. Soc. Psychol.* 13 (1980), 39–80.

[21] Edward L. Deci and Richard M. Ryan. 1985. *Intrinsic Motivation and Self-Determination in Human Behavior.* Springer US.

[22] Edward L. Deci and Richard M. Ryan. 2000. The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior. *Psychol. Inquiry* 11, 4 (2000), 227–268.

[23] Edward L. Deci and Richard M. Ryan. 2002. *Handbook of Self-determination Research.* University Rochester Press.

[24] Edward L. Deci and Richard M. Ryan. 2002. Overview of self-determination theory: An organismic dialectical perspective. *Handbook Self-determ. Res.* (2002), 3–33.

[25] Robert F. DeVellis. 2016. *Scale Development: Theory and Applications.* Vol. 26. Sage publications.

[26] Christoph Dybowski and Sigrid Harendza. 2015. Validation of the physician teaching motivation questionnaire (PTMQ). *BMC Med. Edu.* 15, 1 (2015), 1.

[27] Alek Felstiner. 2011. Working the crowd: Employment and labor law in the crowdsourcing industry. *Berkeley J. Employ. Labor Law* (2011), 143–203.

[28] Sara J. Finney and Christine DiStefano. 2006. Non-normal and categorical data in structural equation modeling. *Structural equation modeling: A second course* (2006), 269–314.

[29] GESIS–Leibniz Institute for the Social Sciences. 2019. Retrieved from http://crowdworkers.info.

[30] Claes Fornell and David F. Larcker. 1981. Evaluating structural equation models with unobservable variables and measurement error. *J. Market. Res.* (1981), 39–50.

[31] James R. Foulds, Levi Boyles, Christopher DuBois, Padhraic Smyth, and Max Welling. 2013. Stochastic collapsed variational Bayesian inference for latent Dirichlet allocation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'13)*, Inderjit S. Dhillon, Yehuda Koren, Rayid Ghani, Ted E. Senator, Paul Bradley, Rajesh Parekh, Jingrui He, Robert L. Grossman, and Ramasamy Uthurusamy (Eds.). ACM, 446–454. DOI: https://doi.org/10.1145/2487575.2487697

[32] Robin D. Froman. 2001. Elements to consider in planning the use of factor analysis. *South. Online J. Nurs. Res.* 2, 5 (2001).

[33] Marylène Gagné and Edward L. Deci. 2005. Self-determination theory and work motivation. *J.f Organiz. Behav.* 26, 4 (2005), 331–362.

[34] Marylène Gagné, Jacques Forest, Marie-Hélène Gilbert, Caroline Aubé, Estelle Morin, and Angela Malorni. 2010. The motivation at work scale: Validation evidence in two languages. *Edu. Psychol. Measure.* 70, 4 (2010), 628–646.

[35] Marylène Gagné, Jacques Forest, Maarten Vansteenkiste, Laurence Crevier-Braud, Anja Van den Broeck, Ann Kristin Aspeli, Jenny Bellerose, Charles Benabou, Emanuela Chemolli, Stefan Tomas Güntert, et al. 2015. The multidimensional work motivation scale: Validation evidence in seven languages and nine countries. *Eur. J. Work Organiz. Psychol.* 24, 2 (2015), 178–196.

[36] Darren George and M. Mallery. 2003. Using SPSS for Windows step by step: A simple guide and reference. Allyn & Bacon, Boston, MA.

[37] Frédéric Guay, Alexandre J. S. Morin, David Litalien, Pierre Valois, and Robert J. Vallerand. 2015. Application of exploratory structural equation modeling to evaluate the academic motivation scale. *J. Exper. Edu.* 83, 1 (2015), 51–82.

[38] Neha Gupta, Andy Crabtree, Tom Rodden, David Martin, and Jacki O'Neill. 2014. Understanding Indian crowdworkers. In *Proceedings of the 17th Conference on Computer Supported Cooperative Work.*

[39] Neha Gupta, David B. Martin, Benjamin V. Hanrahan, and Jacki O'Neill. 2014. Turk-life in India. In *Proceedings of the 18th International Conference on Supporting Group Work*, Sean P. Goggins, Isa Jahnke, David W. McDonald, and Pernille Bjørn (Eds.). ACM, 1–11. DOI: https://doi.org/10.1145/2660398.2660403

[40] Joseph F. Hair, Joseph F. Jr., Barry J. Babin, Rolph E. Anderson, and William C. Black. 2018. *Multivariate Data Analysis (8th Ed.).* Cengage, Hampshire, United Kingdom.

[41] Kevin Hewison and Arne L. Kalleberg. 2013. Precarious work and flexibilization in South and Southeast Asia. *Amer. Behav. Sci.* 57, 4 (2013), 395–402.

[42] Daire Hooper, Joseph Coughlan, and Michael Mullen. 2008. Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods* 6 (2008), 53–60.

[43] John L. Horn and J. Jack McArdle. 1992. A practical and theoretical guide to measurement invariance in aging research. *Exper. Aging Res.* 18, 3 (1992), 117–144.

[44] John Joseph Horton and Lydia B. Chilton. 2010. The labor economics of paid crowdsourcing. In *Proceedings of the 11th ACM Conference on Electronic Commerce (EC'10)*, David C. Parkes, Chrysanthos Dellarocas, and Moshe Tennenholtz (Eds.). ACM, 209–218. DOI : https://doi.org/10.1145/1807342.1807376

[45] Mokter Hossain. 2012. Users' motivation to participate in online crowdsourcing platforms. In *Proceedings of the International Conference on Innovation Management and Technology Research (ICIMTR'12)*. IEEE, 310–315.

[46] Joshua L. Howard, Marylène Gagné, Alexandre J. S. Morin, and Jacques Forest. 2016. Using bifactor exploratory structural equation modeling to test for a continuum structure of motivation. *J. Manage.* (2016), 0149206316645653.

[47] Jeff Howe. 2006. Crowdsourcing: A Definition. Retrieved from http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html.

[48] Panagiotis G. Ipeirotis. 2010. Analyzing the Amazon mechanical turk marketplace. *ACM Crossroads* 17, 2 (2010), 16–21. DOI : https://doi.org/10.1145/1869086.1869094

[49] Panagiotis G. Ipeirotis. 2010. Demographics of mechanical turk. *CeDER Working Papers* (2010).

[50] Arne L. Kalleberg. 2009. Precarious work, insecure workers: Employment relations in transition. *Amer. Sociol. Rev.* 74, 1 (2009), 1–22.

[51] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. Worker motivation in crowdsourcing—A study on mechanical turk. In *Proceedings of the 17th Americas Conference on Information Systems: A Renaissance of Information Technology for Sustainability and Global Competitiveness (AMCIS'11)*, Vallabh Sambamurthy and Mohan Tanniru (Eds.). Association for Information Systems. Retrieved from http://aisel.aisnet.org/amcis2011_submissions/340.

[52] Shashank Khanna, Aishwarya Ratan, James Davis, and William Thies. 2010. Evaluating and improving the usability of mechanical turk for low-income workers in India. In *Proceedings of the 1st ACM Annual Symposium on Computing for Development (ACM DEV'10)*, Andrew M. Dearden, Tapan S. Parikh, and Lakshminarayanan Subramanian (Eds.). ACM, 12. DOI : https://doi.org/10.1145/1926180.1926195

[53] Rex B. Kline. 2015. *Principles and Practice of Structural Equation Modeling*. Guilford Publications.

[54] Jon A. Krosnick. 1999. Survey research. *Ann. Rev. Psychol.* 50, 1 (1999), 537–567.

[55] Siou Chew Kuek, Cecilia Paradi-Guilford, Toks Fayomi, Saori Imaizumi, Panos Ipeirotis, Patricia Pina, Manpreet Singh, et al. 2015. *The Global Opportunity in Online Outsourcing*. Technical Report. The World Bank.

[56] Ihno A. Lee and Kristopher J. Preacher. 2013. Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. Retrieved from http://quantpsy.org/corrtest/corrtest3.htm.

[57] Tak Yeon Lee, Casey Dugan, Werner Geyer, Tristan Ratchford, Jamie C. Rasmussen, N. Sadat Shami, and Stela Lupushor. 2013. Experiments on motivational feedback for crowdsourced workers. In *Proceedings of the 7th International Conference on Weblogs and Social Media (ICWSM'13)*. Emre Kiciman, Nicole B. Ellison, Bernie Hogan, Paul Resnick, and Ian Soboroff (Eds.). The AAAI Press. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6118.

[58] David Litalien, Frédéric Guay, and Alexandre J. S. Morin. 2015. Motivation for PhD studies: Scale development and validation. *Learn. Individ. Dif.* 41 (2015), 1–13.

[59] Irene Mandl, Maurizio Curtarelli, Sara Riso, Oscar Vargas, and Elias Gerogiannis. 2015. New forms of employment. *Eurofound Report*.

[60] Herbert W. Marsh, Kit-Tai Hau, and David Grayson. 2005. Goodness of fit in structural equation models. In *Multivariate applications book series. Contemporary Psychometrics: A Festschrift for Roderick P. McDonald*, A. Maydeu-Olivares and J. J. McArdle (Eds.). Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US.

[61] David B. Martin, Benjamin V. Hanrahan, Jacki O'Neill, and Neha Gupta. 2014. Being a turker. In *Proceedings of the Computer Supported Cooperative Work (CSCW'14)*, Susan R. Fussell, Wayne G. Lutters, Meredith Ringel Morris, and Madhu Reddy (Eds.). ACM, 224–235. DOI : https://doi.org/10.1145/2531602.2531663

[62] Roderick P. McDonald. 1999. *Test Theory: A Unified Treatment*. Lawrence Erlbaum Associates Publishers.

[63] Roger E. Millsap and Margarita Olivera-Aguilar. 2012. Investigating measurement invariance using confirmatory factor analysis. In *Handbook of Structural Equation Modeling*, R. Hoyle (Ed.). Guilford Press, New York.

[64] Babak Naderi. 2018. How to measure motivation? In *Motivation of Workers on Microtask Crowdsourcing Platforms*. Springer, 29–44.

[65] Babak Naderi, Ina Wechsung, Tim Polzehl, and Sebastian Möller. 2014. Development and validation of extrinsic motivation scale for crowdsourcing microtask platforms. In *Proceedings of the International ACM Workshop on Crowdsourcing for Multimedia (CrowdMM'14)*, Judith Redi and Mathias Lux (Eds.). ACM, 31–36. DOI : https://doi.org/10.1145/2660114.2660122

[66] Bloomberg News. 2016. Venezuela's Currency is Collapsing on the Black Market Again. Retrieved from http://www.bloomberg.com/news/articles/2016-11-01/venezuela-s-currency-is-collapsing-on-the-black-market-again.

[67] Ailsa G. Niven and David Markland. 2016. Using self-determination theory to understand motivation for walking: Instrument development and model testing using Bayesian structural equation modelling. *Psychol. Sport. Exercise* 23 (2016), 90–100.

[68] Kimberly A. Noels, Luc G. Pelletier, Richard Clément, and Robert J. Vallerand. 2000. Why are you learning a second language? Motivational orientations and self-determination theory. *Lang. Learn.* 50, 1 (2000), 57–85.

[69] Daniel M. Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *J. Exper. Soc. Psychol.* 45, 4 (2009), 867–872.

[70] Jason W. Osborne and Anna B. Costello. 2009. Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Pan-Pacific Manage. Rev.* 12, 2 (2009), 131–146.

[71] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon mechanical turk. *Judg. Decis. Mak.* 5, 5 (2010), 411–419.

[72] European Parliament. 2016. The situation of workers in the Collaborative Economy. Retrieved from http://www.europarl.europa.eu/RegData/etudes/IDAN/2016/587316/IPOL_IDA(2016)587316_EN.pdf.

[73] Lisa Posch, Arnim Bleier, Fabian Flöck, and Markus Strohmaier. 2018. Characterizing the global crowd workforce: A cross-country comparison of crowdworker demographics. arXiv preprint arXiv:1812.05948.

[74] Lisa Posch, Arnim Bleier, Philipp Schaer, and Markus Strohmaier. 2015. The polylingual labeled topic model. In *Proceedings of the 38th Annual German Conference on Advances in Artificial Intelligence (AI'15) (Lecture Notes in Computer Science)*, Steffen Hölldobler, Markus Krötzsch, Rafael Peñaloza, and Sebastian Rudolph (Eds.), Vol. 9324. Springer, 295–301. DOI: https://doi.org/10.1007/978-3-319-24489-1_26

[75] Lisa Posch, Arnim Bleier, and Markus Strohmaier. 2017. Measuring motivations of crowdworkers: The multidimensional crowdworker motivation scale (first version). *CoRR* http://arxiv.org/abs/1702.01661v1 (2017).

[76] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. 2011. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2778.

[77] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: Shifting demographics in mechanical turk. In *Proceedings of the 28th International Conference on Human Factors in Computing Systems (CHI'10)*, Elizabeth D. Mynatt, Don Schoner, Geraldine Fitzpatrick, Scott E. Hudson, W. Keith Edwards, and Tom Rodden (Eds.). ACM, 2863–2872. DOI: https://doi.org/10.1145/1753846.1753873

[78] Yves Rosseel. 2012. lavaan: An R package for structural equation modeling. *J. Stat. Software* 48, 2 (2012), 1–36. Retrieved from http://www.jstatsoft.org/v48/i02/.

[79] Richard M. Ryan and James P. Connell. 1989. Perceived locus of causality and internalization: Examining reasons for acting in two domains. *J. Personal. Soc. Psychol.* 57, 5 (1989), 749.

[80] Richard M. Ryan and Edward L. Deci. 2000. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemp. Edu. Psychol.* 25, 1 (2000), 54–67.

[81] Cristina Sarasua, Elena Simperl, and Natalya Fridman Noy. 2012. CrowdMap: Crowdsourcing ontology alignment with microtasks. In *Proceedings of the 11th International Semantic Web Conference on the Semantic Web (ISWC'12) (Lecture Notes in Computer Science)*, Philippe Cudré-Mauroux, Jeff Heflin, Evren Sirin, Tania Tudorache, Jérôme Euzenat, Manfred Hauswirth, Josiane Xavier Parreira, Jim Hendler, Guus Schreiber, Abraham Bernstein, and Eva Blomqvist (Eds.), Vol. 7649. Springer, 525–541. DOI: https://doi.org/10.1007/978-3-642-35176-1_33

[82] Albert Satorra and Peter M. Bentler. 2010. Ensuring positiveness of the scaled difference chi-square test statistic. *Psychometrika* 75, 2 (2010), 243–248.

[83] semTools Contributors. 2016. *semTools: Useful Tools for Structural Equation Modeling*. Retrieved from http://CRAN.R-project.org/package=semTools.

[84] Kennon M. Sheldon and Andrew J. Elliot. 1998. Not all personal goals are personal: Comparing autonomous and controlled reasons for goals as predictors of effort and attainment. *Personal. Soc. Psychol. Bull.* 24, 5 (1998), 546–557.

[85] Alexander D. Stajkovic and Fred Luthans. 1997. A meta-analysis of the effects of organizational behavior modification on task performance, 1975–95. *Acad. Manage. J.* 40, 5 (1997), 1122–1149.

[86] James H. Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychol. Bull.* 87, 2 (1980), 245.

[87] Holger Steinmetz. 2013. Analyzing observed composite differences across groups: Is partial measurement invariance enough? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences* 9, 1 (2013), 1–12.

[88] Maxime A. Tremblay, Céline M. Blanchard, Sara Taylor, Luc G. Pelletier, and Martin Villeneuve. 2009. Work extrinsic and intrinsic motivation scale: Its value for organizational psychology research. *Can. J. Behav. Sci.* 41, 4 (2009), 213.

[89] Amazon Mechanical Turk. 2018. Amazon Mechanical Turk: Worker Web Site FAQs. Retrieved from http://www.mturk.com/mturk/help?helpPage=worker#how_paid.

[90] Robert J. Vallerand, Luc G. Pelletier, Marc R. Blais, Nathalie M. Briere, Caroline Senecal, and Evelyne F. Vallieres. 1992. The academic motivation scale: A measure of intrinsic, extrinsic, and amotivation in education. *Edu. Psychol. Measure.* 52, 4 (1992), 1003–1017.

[91] Rens Van de Schoot, Peter Lugtig, and Joop Hox. 2012. A checklist for testing measurement invariance. *Eur. J. Dev. Psychol.* 9, 4 (2012), 486–492.
[92] David L. Vannette and Jon A. Krosnick. 2014. A comparison of survey satisficing and mindlessness. *Wiley Blackwell Handbook Mindful.* 1 (2014), 312.
[93] Roger L. Worthington and Tiffany A. Whittaker. 2006. Scale development research: A content analysis and recommendations for best practices. *Counsel. Psychol.* 34, 6 (2006), 806–838.

Received May 2018; revised February 2019; accepted May 2019

4 Conclusions

The advance of microtask crowdsourcing to support otherwise automated systems has created a new global workforce. Around the world, millions of people work on microtasks to solve problems that computers alone cannot yet solve. Integrated into otherwise automated systems, they perform crucial tasks behind the scenes. These workers, an anonymous and undefined crowd, can be accessed on demand via the interfaces of microtask platforms. As workers on microtask platforms are not considered employees, they are not entitled to any benefits such as sick leave, health insurance, retirement benefits, or vacation pay. Being paid cents at a time, their hourly wages often do not exceed a couple of dollars. Nevertheless, this work provides essential income for many of them, and the rise of work on microtask platforms is an international phenomenon that does not exclude high-income countries.

It is important to understand this emerging form of work and the workforce involved in it, not only for researchers and businesses who exploit the potential of this workforce in order to create better systems but also for policy makers who are concerned with the regulation of this type of work. It is a global responsibility to ensure that this new form of anonymous, hyper-flexible, and completely globalized labor does not lead, as Gray and Suri (2019) warn, to the creation of a "new global underclass."

To gain a better understanding of this new global workforce, this thesis has presented a comprehensive analysis of the socio-demographic characteristics and motivations of the international microtask workforce on the platform Figure Eight. Furthermore, this thesis has presented three use cases that demonstrate how human input from microtasks can be used to complement methods for the analysis of unstructured text in different stages of the machine learning process. The contributions made in this thesis advance our understanding of work on microtask platforms, and they provide a basis for future research regarding this emerging form of labor and the workforce involved in it.

The remainder of this concluding chapter is structured as follows. Section 4.1 summarizes the main results and contributions of this thesis, Section 4.2 discusses the implications of this work and describes a number of potential applications, and Section 4.3 discusses the limitations of this thesis and outlines how these limitations open up directions for future work.

4.1 Results and Contributions

The main contributions of this thesis are (1) a set of use cases that demonstrate the use of the microtask workforce for the analysis of large text corpora in different stages of the machine learning process, (2) an analysis of the socio-demographic characteristics of crowdworkers in different countries, (3) a theory-based and cross-nationally applicable instrument for measuring the motivations of crowdworkers, and (4) an international comparison of crowdworkers' motivations to participate in this type of work. In the following, I summarize the results and contributions presented in this thesis and provide answers to the research questions posed in Section 1.4.

Use of the Microtask Workforce. The use cases presented in Section 3.2 of this thesis demonstrate how microtasks can be used in different stages of the machine learning process to complement machine learning methods for the analysis of large text corpora. Table 4.1 gives an overview of the different types of microtasks that we employed in different stages of the machine learning process. In addition to the overall contribution of demonstrating how microtasks can be used to complement automated methods in different stages of the machine learning process, each of the presented publications contains separate, project-specific contributions.

In Section 3.2.1, this thesis presented the Polylingual Labeled Topic Model, a new topic model for corpora consisting of multilingual labeled documents. In this project, we used microtasks in the *model evaluation* stage of the machine learning process. Specifically, crowdworkers evaluated the semantic coherence of the topics estimated by the new model as well as, for comparison, the semantic coherence of the topics estimated by three existing topic models. The results of our evaluation demonstrated that the new topic model produced topics with a high semantic coherence, while also achieving a good predictive performance. Furthermore, in Posch, Schaer, et al. (2016) we presented a visualization system for displaying probabilistic links that the new topic model is able to create between terms in a thesaurus and classes in a classification system.

Table 4.1: **Microtasks employed in different stages of the machine learning process.** This table gives an overview of how the use cases presented in this thesis employed microtasks to complement automated methods in different stages of the machine learning process.

Stage of the ML process	Microtasks
Data collection	Collect social media accounts of German news media outlets
Data preparation and preprocessing	Identify movie titles, extract relevant keywords/information from text, identify the sentiment associated with movies and keywords
Model evaluation	Word intrusion tasks for the topics estimated by different topic models
Model interpretation	Judge the degree to which topics estimated by a topic model correspond to populist political communication

In the second project, presented in Section 3.2.2, we analyzed online communication by German political actors. The analyses shed light on which topics different German political parties and the movement Pegida addressed in their Facebook posts following the opening of the Pegida Dresden account on 29 December 2014. The project employed microtasks in the *model interpretation* step of the machine learning process. Specifically, crowdworkers interpreted the parameters of a topic model in the context of populist communication, enabling us to analyze the degree of perceived populism exhibited by the different political parties and Pegida over time. The results of our analysis showed that Pegida and the party AfD had a high similarity in topic distribution and that they emphasized populist topics more than other parties.

Section 3.2.3 presented the third use case, an evaluation of the capability of existing recommender algorithms to incorporate information found in narrative descriptions of users' preferences. In this project, we employed microtasks in the *data preparation and preprocessing* stage of the machine learning process. Specifically, we implemented a range of microtasks to extract different types of structured information from the unstructured text of the narrative descriptions. We then compiled this structured information into a reference evaluation dataset, which we used for the evaluation of five different recommender algorithms. The results of this evaluation indicated that by using post-filtering techniques, information extracted from narrative descriptions of users' preferences can help to greatly improve recommendations, provided that the post-filters are carefully configured.

Additionally, in Stier, Bleier, Bonart, Mörsheim, et al. (2018b), we demonstrated the use of the microtask workforce in the *data collection* phase of the machine learning process. In this project, crowdworkers collected social media accounts of mainstream as well as alternative German media on Facebook and Twitter. This collection of accounts enabled us to subsequently collect the social media posts of a wide range of German media outlets.

In sum, these use cases have demonstrated ways in which the microtask workforce can be employed to complement methods for the analysis of unstructured text in different stages of the machine learning process, thus providing answers to the first overarching research question *"How can human input from microtasks complement methods for the analysis of large text corpora in different stages of the machine learning process?"*

Socio-Demographic Characteristics of the Microtask Workforce. The article presented in Section 3.3.1 tackled the second research question: *"What are the socio-demographic characteristics of the international microtask workforce, and do these characteristics differ across countries?"* To answer this question, we conducted a large-scale country-level study on the socio-demographic characteristics of the international microtask workforce and on the importance of microtask income for the workers' lives. The results of our analyses demonstrated that there are substantial differences in the distributions of different socio-demographic characteristics between the workforces in different countries. Regarding the importance of microtask income, the study revealed that in most of the countries included in the analysis, a large proportion of workers uses their income from microtasks to pay for basic expenses such as food, rent, or medical care. Furthermore, the results indicated that between two independent samples taken eight months apart, these characteristics were mostly stable at the country level. This analysis constitutes the first large-scale country-level comparison of the socio-demographic characteristics of crowdworkers that goes beyond an analysis of American and Indian workers on the platform MTurk.

Motivations of the Microtask Workforce. The article presented in Section 3.3.2 set out to answer the third research question: *"Why do people choose to participate in the microtask workforce, and do their motivations differ across countries?"* The main contribution of this article is a theory-based and cross-nationally applicable instrument for measuring the motivations of crowdworkers. This instrument is the first SDT-based motivation scale that was developed specifically for the context of work on microtask platforms and that provides a comprehensive representation of the motivational dimensions according to SDT. Moreover, it is the first motivation scale for the microtask context that is validated across multiple countries and income groups.

In addition to the motivation scale, the article presented the first cross-country comparison of crowdworker motivations. The results of this cross-country comparison showed that there are significant differences between the motivational profiles of the workforces in different countries and income groups. For example, the scores for the identified regulation construct, i.e., the behavior being aligned with personal goals and objectives such as lifestyle preferences or career plans, were significantly lower in high-income countries than in middle- and low-income countries. However, we also found important similarities between the motivations of crowdworkers in different countries. For example, the material external regulation construct received the highest score of all motivation types in all countries, indicating that monetary rewards were the most important motivation for crowdworkers in all countries.

4.2 Implications and Potential Applications

The results of the analyses presented in this thesis are relevant for researchers and practitioners seeking to better understand the international microtask workforce and its use in different stages of the machine learning process. Furthermore, the use cases presented in Section 3.2 contain additional contributions that are relevant for researchers and practitioners interested in the analysis of unstructured text. This section discusses the implications of the studies presented in this thesis and proposes some potential applications.

Use Cases. The use cases presented in Section 3.2 have illustrated how microtasks can be employed in different stages of the machine learning process. As shown in Table 4.1, this set of use cases provides researchers and practitioners with practical examples of how microtasks can be employed in different stages of projects involving machine learning methods. In the following, I describe the implications and potential applications of the contributions made by the individual projects that employed the microtask workforce.

The *Polylingual Labeled Topic Model* presented in Section 3.2.1 enables researchers and practitioners to model text corpora consisting of multilingual, labeled documents. While in our case, the PLL-TM was applied in a two-language setting, the model is capable of modeling corpora containing an arbitrary number of languages. Apart from modeling corpora consisting of documents that are present in multiple natural languages, the PLL-TM can be used to model any collection of documents that have been annotated with thesaurus terms and classified according to a classification system. For such collections of documents, the PLL-TM creates probabilistic links between the thesaurus terms and the classes contained in the classification system. The visualization system that we developed based on the model (see Posch, Schaer, et al., 2016) will help human annotators working simultaneously with a thesaurus and a classification system to quickly determine which thesaurus terms are most strongly associated with which class in the classification system.

The analysis of *Facebook use by the right-wing movement Pegida and German political parties*, presented in Section 3.2.2, has implications for political scientists seeking to understand communication by different political actors on social media. Specifically, the analysis helps researchers to understand how different political actors in Germany use Facebook for their communication with the public and to what extent the different political actors discuss topics perceived as populist communication. The results of the analysis may also inform political discourse by contributing evidence regarding the similarity of the distribution of topics addressed by the different parties and Pegida. Finally, the method used for our analysis may be applied by other researchers to study online populist communication in other countries and contexts.

The *evaluation of narrative-driven movie recommendations on Reddit*, presented in Section 3.2.3, contributes a crowdsourced reference evaluation dataset that can help researchers develop and evaluate new recommender algorithms capable of incorporating contextual information found in narrative descriptions of users' preferences. The results of the evaluation of established recommender algorithms inform researchers and practitioners about the extent to which existing methods are capable of incorporating such information for improving the resulting recommendations. Furthermore, the analyses provide insights into which types of additional information contribute most to improving the results of different algorithms. By using the proposed method, platforms such as IMDb may implement additional functionality that allows users to provide contextual information in order to improve the recommendations they receive.

Socio-Demographic Characteristics of the Microtask Workforce. The results obtained from the analysis of the socio-demographic characteristics and of the importance of microtask income for workers, presented in Section 3.3.1, have implications for task requesters from academia and industry, as well as for microtask platform designers and policy makers. Researchers using microtasks in their projects as well as task requesters from industry should be aware of the fact that the microtask workforces in different countries do not only have varying cultural backgrounds, but that they may also have vastly differing distributions of other characteristics such as gender or education level. When a microtask is offered to workers in all countries, the results may therefore contain different kinds of biases depending on the country distribution of the workers. Task requesters should keep this in mind especially when comparing the results of tasks that were worked on by varying proportions of workers from different countries. The results of our analyses may also inform task requesters about which demographic groups they are likely to be targeting by offering a microtask to workers in a specific country.

The results regarding the importance of microtask income have implications for platform design choices and decisions regarding microtask implementations. Platform designers and task requesters should keep in mind that a large proportion of workers may rely on the income from microtasks to help pay for basic expenses such as food or rent, as the results of our survey indicate. When designing platform functionality such as processes for suspending worker accounts, platform designers should therefore make sure that these processes are transparent and that they offer a possibility for workers to appeal unfair decisions. For task requesters, it is important to ensure that their quality control mechanisms are fair, that task instructions are as clear as possible, and that their task implementations are well-tested. Failing to do so may result in workers not being paid for completed work, in many cases threatening their livelihoods.

For policy makers, the results of the analyses regarding socio-demographic characteristics can inform discussions by providing information on which demographic groups any policies and regulations regarding work on microtask platforms are likely to concern. The results of our analyses regarding the importance of microtask income for workers provide additional input for discussions around new policies and regulations. Finally, the survey developed for this analysis can be used by researchers to collect data on socio-demographic characteristics and the importance of microtask income for workers on other platforms and in additional countries.

Motivations of the Microtask Workforce. The Multidimensional Crowdworker Motivation Scale (MCMS) presented in Section 3.3.2 may be applied by researchers in order to address substantive questions regarding work on microtask platforms. To answer such questions, the six motivational dimensions measured by the scale can serve as an outcome or as an explanatory variable. Potential applications include, among others, predicting worker retention, investigating the relationship between worker motivation and performance on different types of tasks, and conducting comparison studies of the motivations of different crowdworker populations. The instrument can be easily included as a module in the design of microtasks if researchers wish to measure motivation in addition to other variables. Along with the scale, the article presented in Section 3.3.2 also includes instructions for the use of the MCMS.

Platform designers can utilize the scale to analyze whether changes to their platform design or functionality affect the motivations of their workforce and to assess whether changes made to the platform lead to desirable or undesirable changes in motivation. Finally, the analysis of crowdworker motivations presented in this thesis has implications for policy discussions, as it contributes to answering the question as to where in the employment space work on microtask platforms should be located.

4.3 Limitations and Future Work

In the following, I describe the limitations of the work presented in this thesis and outline how these limitations open up directions for future research. The discussion of limitations and potential future work contained in this section is intended as a general overview. A discussion of further, more project-specific limitations and suggestions for future work can be found in the individual publications.

Towards a Comprehensive Taxonomy of Microtasks from the Perspective of the Machine Learning Process. This thesis demonstrated the use of the microtask workforce in different stages of the machine learning process by presenting exemplary use cases. Focusing on the analysis of unstructured text, we employed microtasks in machine learning projects from different domains. In each of these projects, human input from microtasks complemented automated methods in a different stage of the machine learning process, and in each project, crowdworkers were crucial for answering the respective project-specific research questions. However, the uses for microtasks presented in this thesis are not intended as an exhaustive classification of all ways in which microtasks can be employed in different stages of the machine learning process.

In future work, researchers may develop a comprehensive taxonomy of microtasks from the perspective of their uses in different stages of the machine learning process. Such a taxonomy will be useful as a guide for researchers and practitioners seeking to harness the potential of microtasks in their machine learning projects. The uses of microtasks presented in this thesis, summarized in Table 4.1, provide a starting point for developing such a taxonomy.

Extend the Scope of the Analyses Regarding Workforce Characteristics. The analyses presented in this thesis provide new insights into the socio-demographic characteristics and motivations of the international microtask workforce. Nevertheless, to gain a complete picture of the entire global microtask workforce, it will be necessary to study the workforce on other platforms as well as in additional countries. Furthermore, cross-platform comparisons of crowdworker motivations will provide further insights into the similarities and differences of the reasons why people choose to participate in this type of work.

Additional platforms. The studies presented in this thesis have mostly focused on the microtask platform Figure Eight. The reason for this choice was that Figure Eight is the largest microtask platform that also attracts a geographically diverse workforce. Regarding the socio-demographic characteristics, this thesis has presented the first large-scale analysis of socio-demographic characteristics of workers on Figure Eight. The demographics of MTurk's workforce, consisting mainly of workers from the USA and India, are known from previous work. Future work is therefore encouraged to further study the socio-demographic characteristics of workers on platforms other than Figure Eight and MTurk, of which there is still little knowledge, especially at the country level. Regarding the motivations of the microtask workforce, this thesis has presented a theory-based measurement instrument and the first large-scale analysis of crowdworker motivations in different countries. Future research is encouraged to validate the MCMS on additional microtask platforms. A first step towards evaluating the applicability of the MCMS on different platforms has been presented in Section 3.3.2, by conducting a preliminary validation of the motivation scale on the platform MTurk. The results of this first analysis were promising and suggested that the scale is valid for measuring workers' motivations on MTurk. It is therefore likely that the MCMS will constitute a suitable tool for researchers to study the motivations of crowdworkers on other platforms as well.

Additional countries. The analyses presented in this thesis have focused on ten countries. These countries were chosen to reflect a broad cultural diversity as well as income diversity, and they exhibited a high activity on the platform Figure Eight. However, these ten countries are not the only countries that have active workforces on microtask platforms, and if this type of work becomes more widespread, more countries may develop larger microtask workforces. Future work is therefore encouraged to analyze the socio-demographic characteristics and motivations of crowdworkers in additional countries. As the MCMS has already shown to be applicable in ten countries from diverse cultural backgrounds and to produce comparable measurements, it is likely that the scale will be a valid instrument to measure and compare the motivations of crowdworkers in other countries as well.

Cross-Platform Comparisons of Crowdworker Motivations. Conducting comparisons between the motivational profiles of workers on different platforms would contribute further to a more complete understanding of the international microtask workforce. A first analysis regarding the cross-platform comparability of motivations measured with the MCMS has been presented in Section 3.3.2. As microtask platforms are similar in their essential functionality, it is likely that motivations measured with the MCMS are comparable across other platforms as well. Future work is therefore encouraged to further investigate the cross-platform comparability of the MCMS and, provided that the measurements are comparable, use it as a tool to conduct cross-platform comparisons.

Investigate Different Antecedents and Consequences of Crowdworker Motivations. A further direction for future research is the analysis of different antecedents and consequences of crowdworker motivations. While the work presented in this thesis contains preliminary analyses of the relationship between crowdworker motivations and other variables such as time spent on the platform and measures of effort, the focus of the studies regarding crowdworker motivation has been on the development of the MCMS and on the measurement of motivations in different countries. Future research is encouraged to study which factors lead to different motivations in the microtask context and to analyze the consequences of different types of motivation in this context. For example, researchers could investigate how different platforms' design choices affect motivation, or how different motivational profiles are related to outcomes such as the well-being of workers, their performance on different types of tasks, or intentions to stop engaging in this type of work. The MCMS presented in this thesis provides researchers with a tool for conducting such analyses, and it could be integrated in future studies as part of a comprehensive model of work motivation for the microtask context that includes different antecedents and outcomes of motivation. In the following, I describe two specific directions for future research regarding crowdworker motivations that would further advance our understanding of work on microtask platforms.

Analyze the Relationship between Crowdworker Motivations and Socio-Demographic Characteristics. In the article presented in Section 3.3.2, we analyzed the differences in motivations of crowdworkers located in different countries. First analyses regarding other variables also showed that crowdworkers who spend a lot of time on the platform tend to have different motivations than workers who spend little time on the platform. Future research is encouraged to further study the relationship between different socio-demographic characteristics of workers and their motivations, going beyond the country of residence as an indicator of difference. For example, factors such as household income, education, the workers' employment status aside from their work on microtask platforms, or the workers' degree of dependency on income from microtasks might be related to different motivational profiles.

Analyze the Relationship between Crowdworker Motivations and Task Performance. The MCMS could be used in future work to investigate the effect of different types of motivation on workers' performance on different types of tasks. Work motivation has been shown to play an important role in job performance in other contexts (see, e.g., C. A. O'Reilly and Chatman, 1994; Van Knippenberg, 2000; Pinder, 2014; Deci, Olafsen, et al., 2017). However, it is not the only factor that determines the level of a worker's performance, and future research investigating the relationship between the motivations of crowdworkers and their performance on tasks should take into account several issues.

As Pinder (2014) notes, it is a mistake to automatically assume that poor job performance is the result of low motivation, when, in reality, the problem might stem from many other factors. Besides external factors such as the physical environment of the workplace, the ability of the worker is an important factor in job performance (Pinder, 2014). A number of studies (see, e.g., Borman et al., 1991; C. A. O'Reilly and Chatman, 1994; Hirschfeld et al., 2004) have shown that ability and motivation are both necessary for performance and that there is likely to be an interaction effect between motivation and ability. Therefore, in studies investigating the effect of crowdworkers' motivation on their task performance, other factors such as ability should be taken into account in addition to motivation.

Moreover, future studies that analyze task performance of crowdworkers as an outcome variable should take great care in defining their measures of performance. Measured performance heavily depends on decisions such as who sets the goals, how many goals are in place simultaneously, whether these goals are mutually exclusive, as well as how, when, and by whom the performance is measured, and whether the measure used is absolute or relative to other people (see, e.g., Mitchell and C. O'Reilly, 1983; Pinder, 2014). Nevertheless, a careful analysis of the relationship between crowdworker motivation and their performance on different types of microtasks has the potential to provide important insights for researchers and practitioners using microtasks in their projects, and it may provide guidance for microtask platform designers who want to improve their platforms.

This thesis has set out to deepen our understanding of work on microtask platforms and of the international workforce involved in this type of work. With the contributions made in this thesis, I have provided a clearer picture of the socio-demographic characteristics and motivations of the international microtask workforce. Furthermore, I have showcased how the microtask workforce can be employed in different stages of the machine learning process for the analysis of large text corpora. These contributions provide a foundation for further research regarding this emerging form of work and the workforce involved in it.

Bibliography

Amazon Mechanical Turk (2015). *Bringing future innovation to Amazon Mechanical Turk*. https://blog.mturk.com/bringing-future-innovation-to-mechanical-turk-c67e489e0c37. Accessed: 2020-05-11 (cit. on pp. 5, 52).

Amazon Mechanical Turk (2016). *Celebrating 11 years of artificial, artificial intelligence*. https://blog.mturk.com/celebrating-11-years-of-artificial-artificial-intelligence-e94ec6a56b0b. Accessed: 2020-02-20 (cit. on p. 4).

Amazon Mechanical Turk (2018). *Amazon Mechanical Turk: Worker web site FAQs*. https://www.mturk.com/mturk/help?helpPage=worker#how_paid. Accessed: 2019-04-15 (cit. on pp. 6, 53).

Amazon Mechanical Turk (2019). https://blog.mturk.com/amazon-mechanical-turk-workers-in-23-countries-outside-of-the-us-can-now-transfer-their-earnings-98ec29ef7f7f. Accessed: 2020-03-30 (cit. on p. 7).

Amazon Mechanical Turk (2020). *Participation agreement*. https://www.mturk.com/participation-agreement. Accessed: 2020-01-30 (cit. on p. 9).

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., and Zimmermann, T. (2019). "Software engineering for machine learning: A case study." In: *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. IEEE, pp. 291–300 (cit. on p. 25).

Antin, J. and Shaw, A. D. (2012). "Social desirability bias and self-reports of motivation: A study of Amazon Mechanical Turk in the US and India." In: *CHI Conference on Human Factors in Computing Systems, CHI '12, Austin, TX, USA - May 05 - 10, 2012*. Ed. by Konstan, J. A., Chi, E. H., and Höök, K. ACM, pp. 2925–2934. ISBN: 978-1-4503-1015-4. DOI: `10.1145/2207676.2208699`. URL: `http://doi.acm.org/10.1145/2207676.2208699` (cit. on p. 58).

Appen (2019). *Appen completes acquisition of Figure Eight and achieves critical integration milestone*. `https://appen.com/press-release/appen-completes-acquisition-of-figure-eight-and-achieves-critical-integration-milestone/`. Accessed: 2020-03-20 (cit. on p. 6).

Barber, D. (2012). *Bayesian reasoning and machine learning*. Cambridge University Press (cit. on pp. 29, 30).

Barbosa, N. M. and Chen, M. (2019). "Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–12 (cit. on p. 8).

Barret, V. (2009). *Dolores Labs vets web sites on the cheap*. `https://www.forbes.com/forbes/2009/0330/048-you-know-it.html`. Accessed: 2019-12-20 (cit. on p. 5).

Bastian, B. and Haslam, N. (2011). "Experiencing dehumanization: Cognitive and emotional effects of everyday dehumanization." In: *Basic and Applied Social Psychology* 33.4, pp. 295–303 (cit. on p. 8).

Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., and Mikhaylov, S. (2016). "Crowd-sourced text analysis: Reproducible and agile production of political data." In: *American Political Science Review* 110.2, pp. 278–295 (cit. on p. 49).

Bentler, P. M. and Bonett, D. G. (1980). "Significance tests and goodness of fit in the analysis of covariance structures." In: *Psychological Bulletin* 88.3, p. 588 (cit. on p. 39).

Berg, J. (2015). "Income security in the on-demand economy: Findings and policy lessons from a survey of crowdworkers." In: *Comparative Labor Law & Policy Journal* 37, p. 543 (cit. on pp. 9–11, 17, 52, 54).

Berg, J., Furrer, M., Harmon, E., Rani, U., and Silberman, M. S. (2018). *Digital labour platforms and the future of work: towards decent work in the online world*. International Labour Office Geneva (cit. on pp. 10, 55).

Berinsky, A. J., Huber, G. A., and Lenz, G. S. (2012). "Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk." In: *Political Analysis* 20.3, pp. 351–368 (cit. on p. 53).

Bernstein, M. S., Little, G., Miller, R. C., Hartmann, B., Ackerman, M. S., Karger, D. R., Crowell, D., and Panovich, K. (2010). "Soylent: a word processor with a crowd inside." In: *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*. UIST '10. New York, New York, USA: Association for Computing Machinery, pp. 313–322. ISBN: 9781450302715. DOI: 10.1145/1866029.1866078. URL: https://doi.org/10.1145/1866029.1866078 (cit. on p. 50).

Berry, D. M. (2019). "Against infrasomatization: Towards a critical theory of algorithms." In: *Data Politics*. Routledge, pp. 43–63 (cit. on p. 8).

Bertschek, I., Ohnemus, J., and Viete, S. (2015). *Befragung zum sozioökonomischen Hintergrund und zu den Motiven von Crowdworkern an das Bundesministerium für Arbeit und Soziales, Berlin: Endbericht zur Kurzexpertise*. Tech. rep. ZEW-Gutachten und Forschungsberichte (cit. on p. 55).

Bezos, J. (2006). *Opening keynote and keynote interview with Jeff Bezos*. https://techtv.mit.edu/videos/16180-opening-keynote-and-keynote-interview-with-jeff-bezos. Accessed: 2019-12-20 (cit. on pp. 4, 8).

Blais, M., Lachance, L., Vallerand, R., Briere, N., and Riddle, A. (1993). "The work motivation inventory." In: *Revue Quebecoise de Psychologie* 14, pp. 185–215 (cit. on p. 45).

Blei, D. M. (2012). "Probabilistic topic models." In: *Communications of the ACM* 55, pp. 77–84 (cit. on p. 29).

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2002). "Latent Dirichlet allocation." In: *Advances in Neural Information Processing Systems 14*. Ed. by Dietterich, T. G., Becker, S., and Ghahramani, Z. MIT Press, pp. 601–608. URL: http://papers.nips.cc/paper/2070-latent-dirichlet-allocation.pdf (cit. on p. 30).

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet allocation." In: *Journal of Machine Learning Research* 3, pp. 993–1022. URL: `http://www.jmlr.org/papers/v3/blei03a.html` (cit. on p. 29).

Bleier, A. (2012). "A simple non-parametric topic mixture for authors and documents." In: *arXiv preprint arXiv:1211.6248* (cit. on p. 33).

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons (cit. on pp. 34, 37).

Bontcheva, K., Derczynski, L., and Roberts, I. (2017). "Crowdsourcing named entity recognition and entity linking corpora." In: *Handbook of Linguistic Annotation*. Springer, pp. 875–892 (cit. on p. 49).

Borman, W. C., White, L. A., Pulakos, E. D., and Oppler, S. H. (1991). "Models of supervisory job performance ratings." In: *Journal of Applied Psychology* 76.6, p. 863 (cit. on p. 192).

Borromeo, R. M. and Toyama, M. (2015). "Automatic vs. crowdsourced sentiment analysis." In: *Proceedings of the 19th International Database Engineering & Applications Symposium*. ACM, pp. 90–95 (cit. on p. 48).

Brawley, A. M. and Pury, C. L. (2016). "Work experiences on MTurk: Job satisfaction, turnover, and information sharing." In: *Computers in Human Behavior* 54, pp. 531–546 (cit. on p. 58).

Bu, Q., Simperl, E., Chapman, A., and Maddalena, E. (2019). "Quality assessment in crowdsourced classification tasks." In: *International Journal of Crowd Science* (cit. on p. 48).

Buhrmester, M., Kwang, T., and Gosling, S. D. (2011). "Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data?" In: *Perspectives on Psychological Science* 6.1, pp. 3–5 (cit. on pp. 53, 57).

Byrne, B. M., Shavelson, R. J., and Muthén, B. (1989). "Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance." In: *Psychological Bulletin* 105.3, p. 456 (cit. on p. 41).

Caines, A., Bentz, C., Graham, C., Polzehl, T., and Buttery, P. (2016). "Crowdsourcing a multilingual speech corpus: Recording, transcription and annotation of the crowded corpus." In: *Proceedings of LREC, Portoroz, Slovenia*, pp. 23–30 (cit. on p. 48).

Callison-Burch, C. (2009). "Fast, cheap, and creative: Evaluating translation quality using Amazon's Mechanical Turk." In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*. Association for Computational Linguistics, pp. 286–295 (cit. on p. 50).

Chandler, D. and Kapelner, A. (2013). "Breaking monotony with meaning: Motivation in crowdsourcing markets." In: *Journal of Economic Behavior & Organization* 90, pp. 123–133 (cit. on p. 60).

Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C., and Blei, D. M. (2009). "Reading tea leaves: How humans interpret topic models." In: *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009.* Ed. by Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A. Curran Associates, Inc., pp. 288–296. ISBN: 9781615679119. URL: http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models (cit. on p. 50).

Chen, F. F. (2007). "Sensitivity of goodness of fit indexes to lack of measurement invariance." In: *Structural Equation Modeling* 14.3, pp. 464–504 (cit. on p. 40).

Chen, W.-C., Suri, S., and Gray, M. L. (2019). "More than money: Correlation among worker demographics, motivations, and participation in online labor market." In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 13. 01, pp. 134–145 (cit. on p. 59).

Cherry, M. A. (2016). "Beyond misclassification: The digital transformation of work." In: *Comparative Labor Law & Policy Journal* 37, p. 577 (cit. on pp. 8, 9).

Cheung, G. W. and Rensvold, R. B. (2000). "Assessing extreme and acquiescence response sets in cross-cultural research using structural equations modeling." In: *Journal of Cross-Cultural Psychology* 31.2, pp. 187–212 (cit. on p. 40).

Cheung, G. W. and Rensvold, R. B. (2002). "Evaluating goodness-of-fit indexes for testing measurement invariance." In: *Structural Equation Modeling* 9.2, pp. 233–255 (cit. on p. 40).

Codagnone, C., Abadie, F., and Biagi, F. (2016). *The future of work in the 'sharing economy'. Market efficiency and equitable opportunities or unfair precarisation?* Tech. rep. Institute for Prospective Technological Studies, Science for Policy report by the Joint Research Centre (cit. on p. 10).

Costello, A. B. and Osborne, J. (2005). "Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis." In: *Practical Assessment, Research, and Evaluation* 10.1, p. 7 (cit. on p. 34).

De Stefano, V. (2016). "The rise of the just-in-time workforce: On-demand work, crowdwork, and labor protection in the gig-economy." In: *Comparative Labor Law & Policy Journal* 37, p. 471 (cit. on pp. 7, 8, 10).

Deci, E. L. (1971). "Effects of externally mediated rewards on intrinsic motivation." In: *Journal of Personality and Social Psychology* 18.1, p. 105 (cit. on p. 44).

Deci, E. L., Koestner, R., and Ryan, R. M. (1999). "A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation." In: *Psychological Bulletin* 125.6, p. 627 (cit. on p. 44).

Deci, E. L., Olafsen, A. H., and Ryan, R. M. (2017). "Self-determination theory in work organizations: The state of a science." In: *Annual Review of Organizational Psychology and Organizational Behavior* 4, pp. 19–43 (cit. on pp. 41, 45, 192).

Deci, E. L. and Ryan, R. M. (1980). "The empirical exploration of intrinsic motivational processes." In: *Advances in Experimental Social Psychology* 13, pp. 39–80 (cit. on pp. 41, 44).

Deci, E. L. and Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Springer US. ISBN: 978-1-4899-2273-1 (cit. on pp. 41, 44, 45).

Deci, E. L. and Ryan, R. M. (2000). "The "what" and "why" of goal pursuits: Human needs and the self-determination of behavior." In: *Psychological Inquiry* 11.4, pp. 227–268 (cit. on pp. 41–43, 45).

Deci, E. L. and Ryan, R. M. (2002). *Handbook of self-determination research*. University Rochester Press (cit. on pp. 41–43).

Demartini, G., Difallah, D. E., and Cudré-Mauroux, P. (2012). "ZenCrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking." In: *Proceedings of the 21st International Conference on World Wide Web*. ACM, pp. 469–478 (cit. on p. 49).

Demartini, G., Difallah, D. E., Gadiraju, U., and Catasta, M. (2017). "An introduction to hybrid human-machine information systems." In: *Foundations and Trends in Web Science* 7.1, pp. 1–87. DOI: 10.1561/1800000025. URL: https://doi.org/10.1561/1800000025 (cit. on p. 47).

Deng, X. N. and Joshi, K. (2016). "Why individuals participate in micro-task crowdsourcing work environment: Revealing crowdworkers' perceptions." In: *Journal of the Association for Information Systems* 17.10, p. 3 (cit. on pp. 10, 60).

Dietz, L., Bickel, S., and Scheffer, T. (2007). "Unsupervised prediction of citation influences." In: *Proceedings of the 24th International Conference on Machine Learning*. ACM, pp. 233–240 (cit. on p. 33).

Difallah, D., Filatova, E., and Ipeirotis, P. (2018). "Demographics and dynamics of Mechanical Turk workers." In: *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*. ACM, pp. 135–143 (cit. on p. 52).

Eberhard, L., Walk, S., Posch, L., and Helic, D. (2019). "Evaluating narrative-driven movie recommendations on Reddit." In: *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI 2019, Marina del Ray, CA, USA, March 17-20, 2019*. Ed. by Fu, W., Pan, S., Brdiczka, O., Chau, P., and Calvary, G. ACM, pp. 1–11. DOI: 10.1145/3301275.3302287. URL: https://doi.org/10.1145/3301275.3302287 (cit. on pp. 4, 14–16).

European Commission (2016). *A european agenda for the collaborative economy*. http://ec.europa.eu/DocsRoom/documents/16881. Accessed: 2020-05-11 (cit. on p. 10).

European Parliament (2016). *The situation of workers in the collaborative economy*. http://www.europarl.europa.eu/RegData/etudes/IDAN/2016/587316/IPOL_IDA(2016)587316_EN.pdf. Accessed: 2020-05-11 (cit. on pp. 2, 10, 11).

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J. (1999). "Evaluating the use of exploratory factor analysis in psychological research." In: *Psychological Methods* 4.3, p. 272 (cit. on pp. 35, 36).

Fayyad, U. M., Piatetsky-Shapiro, G., and Smyth, P. (1996). "Knowledge discovery and data mining: towards a unifying framework." In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon, USA*. Ed. by Simoudis, E., Han, J., and Fayyad, U. M. AAAI Press, pp. 82–88. URL: http://www.aaai.org/Library/KDD/1996/kdd96-014.php (cit. on p. 25).

Felstiner, A. (2011). "Working the crowd: Employment and labor law in the crowdsourcing industry." In: *Berkeley Journal of Employment and Labor Law*, pp. 143–203 (cit. on pp. 8–10).

Feyisetan, O., Luczak-Roesch, M., Simperl, E., Tinati, R., and Shadbolt, N. (2015). "Towards hybrid NER: A study of content and crowdsourcing-related performance factors." In: *European Semantic Web Conference*. Springer, pp. 525–540 (cit. on p. 49).

Feyisetan, O. and Simperl, E. (2019). "Beyond monetary incentives: Experiments in paid microtask contests." In: *ACM Transactions on Social Computing* 2.2, p. 6 (cit. on p. 61).

Fieseler, C., Bucher, E., and Hoffmann, C. P. (2019). "Unfairness by design? The perceived fairness of digital labor on crowdworking platforms." In: *Journal of Business Ethics* 156.4, pp. 987–1005 (cit. on p. 9).

Figure Eight (2018a). *Figure Eight master terms of service*. https://www.figure-eight.com/legal/. Accessed: 2020-01-30 (cit. on p. 9).

Figure Eight (2018b). *Focused on the future with a new name*. https://www.figure-eight.com/focused-future-new-name/. Accessed: 2019-12-20 (cit. on p. 6).

Finin, T., Murnane, W., Karandikar, A., Keller, N., Martineau, J., and Dredze, M. (2010). "Annotating named entities in Twitter data with crowdsourcing." In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, pp. 80–88 (cit. on p. 49).

Finkin, M. (2016). "Beclouded work in historical perspective." In: *Comparative Labor Law & Policy Journal* 37.3, pp. 16–12 (cit. on p. 8).

Gadiraju, U., Fetahu, B., and Kawase, R. (2015). "Training workers for improving performance in crowdsourcing microtasks." In: *Design for Teaching and Learning in a Networked World*. Springer, pp. 100–114 (cit. on p. 48).

Gadiraju, U., Kawase, R., and Dietze, S. (2014). "A taxonomy of microtasks on the web." In: *Proceedings of the 25th ACM Conference on Hypertext and Social Media*. ACM, pp. 218–223 (cit. on p. 47).

Gagné, M. and Deci, E. L. (2005). "Self-determination theory and work motivation." In: *Journal of Organizational Behavior* 26.4, pp. 331–362 (cit. on pp. 42–45).

Gagné, M., Forest, J., Gilbert, M.-H., Aubé, C., Morin, E., and Malorni, A. (2010). "The Motivation at Work Scale: Validation evidence in two languages." In: *Educational and Psychological Measurement* 70.4, pp. 628–646 (cit. on p. 45).

Gagné, M., Forest, J., Vansteenkiste, M., Crevier-Braud, L., Van den Broeck, A., Aspeli, A. K., Bellerose, J., Benabou, C., Chemolli, E., Güntert, S. T., et al. (2015). "The Multidimensional Work Motivation Scale: Validation evidence in seven languages and nine countries." In: *European Journal of Work and Organizational Psychology* 24.2, pp. 178–196 (cit. on p. 45).

Geiger, D., Rosemann, M., Fielt, E., and Schader, M. (2012). "Crowdsourcing information systems-definition typology, and design." In: *Proceedings of the International Conference on Information Systems, ICIS 2012*. AISeL (cit. on p. 47).

Geiger, D., Seedorf, S., Schulze, T., Nickerson, R. C., and Schader, M. (2011). "Managing the crowd: Towards a taxonomy of crowdsourcing processes." In: *A Renaissance of Information Technology for Sustainability and Global Competitiveness. 17th Americas Conference on Information Systems, AMCIS 2011, Detroit, Michigan, USA, August 4-8 2011*. Ed. by Sambamurthy, V. and Tanniru, M. Association for Information Systems. URL: `http://aisel.aisnet.org/amcis2011%5C_submissions/430` (cit. on p. 46).

Geman, S. and Geman, D. (1984). "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6, pp. 721–741 (cit. on p. 31).

Goodman, J. K. and Paolacci, G. (2017). "Crowdsourcing consumer research." In: *Journal of Consumer Research* 44.1, pp. 196–210 (cit. on p. 52).

Gray, M. L. and Suri, S. (2019). *Ghost work: How to stop Silicon Valley from building a new global underclass*. Eamon Dolan Books (cit. on pp. 2, 8–10, 179).

Griffiths, T. L. and Steyvers, M. (2004). "Finding scientific topics." In: *Proceedings of the National Academy of Sciences* (cit. on p. 28).

Guo, L., Mays, K., Lai, S., Jalal, M., Ishwar, P., and Betke, M. (2019). "Accurate, fast, but not always cheap: Evaluating "crowdcoding" as an alternative approach to analyze social media data." In: *Journalism & Mass Communication Quarterly* (cit. on p. 48).

Gupta, N., Crabtree, A., Rodden, T., Martin, D., and O'Neill, J. (2014). "Understanding Indian crowdworkers." In: *Proceedings of the 17th Conference on Computer Supported Cooperative Work* (cit. on pp. 10, 60).

Gupta, N., Martin, D. B., Hanrahan, B. V., and O'Neill, J. (2014). "Turklife in India." In: *Proceedings of the 18th International Conference on Supporting Group Work, Sanibel Island, FL, USA, November 09 - 12, 2014*. Ed. by Goggins, S. P., Jahnke, I., McDonald, D. W., and Bjørn, P. ACM, pp. 1–11. ISBN: 978-1-4503-3043-5. DOI: 10.1145/2660398.2660403. URL: http://doi.acm.org/10.1145/2660398.2660403 (cit. on p. 60).

Ha, A. (2012). *CrowdFlower co-founder Lukas Biewald becomes CEO (again)*. https://techcrunch.com/2012/03/22/crowdflower-lukas-biewald-returns/. Accessed: 2019-12-20 (cit. on p. 6).

Hackman, J. R. and Oldham, G. R. (1975). "Development of the job diagnostic survey." In: *Journal of Applied Psychology* 60.2, p. 159 (cit. on p. 60).

Hackman, J. R. and Oldham, G. R. (1980). *Work redesign*. Addison-Wesley (cit. on p. 57).

Hair Joseph F. Jr., J. F., Babin, B. J., Anderson, R. E., and Black, W. C. (2018). *Multivariate data analysis (8th ed.)* Hampshire, United Kingdom: Cengage. ISBN: 9781473756540 (cit. on pp. 27, 28, 34–39).

Harman, H. H. (1976). *Modern factor analysis*. University of Chicago press (cit. on pp. 34, 35).

Haselmayer, M. and Jenny, M. (2017). "Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding." In: *Quality & Quantity* 51.6, pp. 2623–2646 (cit. on p. 48).

Haslam, N. (2006). "Dehumanization: An integrative review." In: *Personality and Social Psychology Review* 10.3, pp. 252–264 (cit. on p. 8).

Heinrich, G. (2005). *Parameter estimation for text analysis*. Tech. rep. Fraunhofer IGD (cit. on p. 29).

Hewison, K. and Kalleberg, A. L. (2013). "Precarious work and flexibilization in South and Southeast Asia." In: *American Behavioral Scientist* 57.4, pp. 395–402 (cit. on pp. 2, 7).

Hindenburg, C. F. (1784). *Ueber den Schachspieler des Herrn von Kempelen: nebst einer Abbildung und Beschreibung seiner Sprachmaschine*. Müler (cit. on p. 5).

Hirschfeld, R. R., Lawson, L., and Mossholder, K. W. (2004). "Moderators of the relationship between cognitive ability and performance: General versus context–specific achievement motivation." In: *Journal of Applied Social Psychology* 34.11, pp. 2389–2409 (cit. on p. 192).

Hirth, M., Hoßfeld, T., and Tran-Gia, P. (2011). "Human cloud as emerging internet application-anatomy of the microworkers crowdsourcing platform." In: *University of Würzburg Institute of Computer Science Research Report Series* (cit. on p. 55).

Hoffmann, L. (2009). "Crowd control." In: *Communications of the ACM* 52.3, pp. 16–17 (cit. on p. 48).

Hooper, D., Coughlan, J., and Mullen, M. R. (2008). "Structural equation modelling: Guidelines for determining model fit." In: *Electronic Journal of Business Research Methods* 6.1, pp. 53–60 (cit. on pp. 38, 39).

Horn, J. L. and McArdle, J. J. (1992). "A practical and theoretical guide to measurement invariance in aging research." In: *Experimental Aging Research* 18.3, pp. 117–144 (cit. on p. 40).

Horton, J. J. and Chilton, L. B. (2010). "The labor economics of paid crowdsourcing." In: *Proceedings of the 11th ACM Conference on Electronic Commerce (EC-2010), Cambridge, Massachusetts, USA, June 7-11, 2010*. Ed. by Parkes, D. C., Dellarocas, C., and Tennenholtz, M. ACM, pp. 209–218. ISBN: 978-1-60558-822-3. DOI: 10.1145/1807342.1807376. URL: http://doi.acm.org/10.1145/1807342.1807376 (cit. on p. 9).

Hossain, M. (2012). "Users' motivation to participate in online crowdsourcing platforms." In: *2012 International Conference on Innovation Management and Technology Research*. IEEE, pp. 310–315 (cit. on p. 58).

Howe, J. (2006). *Crowdsourcing: A definition*. http://crowdsourcing.typepad.com/cs/2006/06/crowdsourcing_a.html. Accessed: 2020-05-11 (cit. on p. 3).

Hoyle, R. H. (2012). *Handbook of structural equation modeling*. Guilford Press (cit. on p. 37).

Hsueh, P.-Y., Melville, P., and Sindhwani, V. (2009). "Data quality from crowdsourcing: A study of annotation selection criteria." In: *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*. Association for Computational Linguistics, pp. 27–35 (cit. on p. 48).

Huff, C. and Tingley, D. (2015). "'Who are these people?' Evaluating the demographic characteristics and political preferences of MTurk survey respondents." In: *Research & Politics* 2.3 (cit. on p. 53).

Ipeirotis, P. G. (2010a). "Demographics of Mechanical Turk." In: *CeDER Working Papers* (cit. on pp. 52, 56).

Ipeirotis, P. G. (2010b). "Analyzing the Amazon Mechanical Turk marketplace." In: *ACM Crossroads* 17.2, pp. 16–21. DOI: 10.1145/1869086.1869094. URL: http://doi.acm.org/10.1145/1869086.1869094 (cit. on p. 52).

Ipeirotis, P. G., Provost, F., and Wang, J. (2010). "Quality management on Amazon Mechanical Turk." In: *Proceedings of the ACM SIGKDD Workshop on Human Computation*. ACM, pp. 64–67 (cit. on p. 48).

Irani, L. (2015a). *Justice for "data janitors"*. https://www.publicbooks.org/justice-for-data-janitors/. Accessed: 2020-05-11 (cit. on p. 8).

Irani, L. (2015b). "The cultural work of microwork." In: *New Media & Society* 17.5, pp. 720–739 (cit. on p. 8).

Jennrich, R. (2007). "Rotation algorithms: From beginning to end." In: *Handbook of Latent Variable and Related Models*. Elsevier, pp. 45–63 (cit. on pp. 35, 36).

Jiang, L., Wagner, C., and Nardi, B. (2015). "Not just in it for the money: A qualitative investigation of workers' perceived benefits of micro-task crowdsourcing." In: *2015 48th Hawaii International Conference on System Sciences*. IEEE, pp. 773–782 (cit. on p. 60).

Jöreskog, K. G. (1970). "A general method for estimating a linear structural equation system." In: *ETS Research Bulletin Series* 1970.2, pp. i–41 (cit. on p. 38).

Jöreskog, K. G. (1971). "Simultaneous factor analysis in several populations." In: *Psychometrika* 36.4, pp. 409–426 (cit. on p. 40).

Jöreskog, K. G. and Goldberger, A. S. (1972). "Factor analysis by generalized least squares." In: *Psychometrika* 37.3, pp. 243–260 (cit. on p. 38).

Kalleberg, A. L. (2009). "Precarious work, insecure workers: Employment relations in transition." In: *American Sociological Review* 74.1, pp. 1–22 (cit. on pp. 2, 7).

Kaplan, D. (2008). *Structural equation modeling: Foundations and extensions*. Vol. 10. Sage Publications (cit. on pp. 34, 37).

Kaplan, T., Saito, S., Hara, K., and Bigham, J. P. (2018). "Striving to earn more: A survey of work strategies and tool use among crowd workers." In: *Sixth AAAI Conference on Human Computation and Crowdsourcing* (cit. on p. 9).

Karger, D. R., Oh, S., and Shah, D. (2011). "Iterative learning for reliable crowdsourcing systems." In: *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011.* Ed. by Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F. C. N., and Weinberger, K. Q., pp. 1953–1961. URL: `http://papers.nips.cc/paper/4396-iterative-learning-for-reliable-crowdsourcing-systems` (cit. on p. 48).

Kaufmann, N., Schulze, T., and Veit, D. (2011). "More than fun and money. Worker motivation in crowdsourcing - A study on Mechanical Turk." In: *A Renaissance of Information Technology for Sustainability and Global Competitiveness. 17th Americas Conference on Information Systems, AMCIS 2011, Detroit, Michigan, USA, August 4-8 2011*. Ed. by Sambamurthy, V. and Tanniru, M. Association for Information Systems. URL: http://aisel.aisnet.org/amcis2011_submissions/340 (cit. on pp. 56–58).

Kessler, S. (2014). *Pixel & dimed: On (not) getting by in the gig economy*. https://www.fastcompany.com/3027355/pixel-and-dimed-on-not-getting-by-in-the-gig-economy. Accessed: 2020-02-20 (cit. on p. 9).

Khanna, S., Ratan, A., Davis, J., and Thies, W. (2010). "Evaluating and improving the usability of Mechanical Turk for low-income workers in India." In: *First ACM Annual Symposium on Computing for Development, ACM DEV '10, London, United Kingdom, December 17 - 18, 2010*. Ed. by Dearden, A. M., Parikh, T. S., and Subramanian, L. ACM, p. 12. ISBN: 978-1-4503-0473-3. DOI: 10.1145/1926180.1926195. URL: http://doi.acm.org/10.1145/1926180.1926195 (cit. on p. 9).

Kittur, A., Nickerson, J. V., Bernstein, M. S., Gerber, E., Shaw, A. D., Zimmerman, J., Lease, M., and Horton, J. J. (2013). "The future of crowd work." In: *Computer Supported Cooperative Work, CSCW 2013, San Antonio, TX, USA, February 23-27, 2013*. Ed. by Bruckman, A., Counts, S., Lampe, C., and Terveen, L. G. ACM, pp. 1301–1318. ISBN: 978-1-4503-1331-5. DOI: 10.1145/2441776.2441923. URL: http://doi.acm.org/10.1145/2441776.2441923 (cit. on p. 8).

Kline, R. B. (2015). *Principles and practice of structural equation modeling*. The Guilford Press (cit. on pp. 27, 34–39).

Kuek, S. C., Paradi-Guilford, C., Fayomi, T., Imaizumi, S., Ipeirotis, P., Pina, P., and Singh, M. (2015). *The global opportunity in online outsourcing*. Tech. rep. The World Bank (cit. on pp. 6, 17, 54).

Lakhani, K. R. and Wolf, R. G. (2005). "Why hackers do what they do: Understanding motivation and effort in free/open source software projects." In: *Perspectives on Free and Open Source Software* (cit. on p. 57).

Lawson, N., Eustice, K., Perkowitz, M., and Yetisgen-Yildiz, M. (2010). "Annotating large email datasets for named entity recognition with Mechanical Turk." In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, pp. 71–79 (cit. on p. 49).

Lee, T. Y., Dugan, C., Geyer, W., Ratchford, T., Rasmussen, J. C., Shami, N. S., and Lupushor, S. (2013). "Experiments on motivational feedback for crowdsourced workers." In: *Proceedings of the Seventh International Conference on Weblogs and Social Media, ICWSM 2013, Cambridge, Massachusetts, USA, July 8-11, 2013*. Ed. by Kiciman, E., Ellison, N. B., Hogan, B., Resnick, P., and Soboroff, I. The AAAI Press. ISBN: 978-1-57735-610-3. URL: `http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/view/6118` (cit. on p. 60).

Lin, J. (1991). "Divergence measures based on the Shannon entropy." In: *IEEE Transactions on Information Theory* 37.1, pp. 145–151 (cit. on p. 18).

Lind, F., Gruber, M., and Boomgaarden, H. G. (2017). "Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs." In: *Communication Methods and Measures* 11.3, pp. 191–209 (cit. on pp. 49, 51).

Litman, L., Robinson, J., and Rosenzweig, C. (2015). "The relationship between motivation, monetary compensation, and data quality among US-and India-based workers on Mechanical Turk." In: *Behavior Research Methods* 47.2, pp. 519–528 (cit. on p. 57).

Mandl, I., Curtarelli, M., Riso, S., Vargas, O., and Gerogiannis, E. (2015). *New forms of employment*. Eurofound Report (cit. on pp. 1, 3).

Marsh, H. W., Hau, K.-T., and Grayson, D. (2005). "Goodness of fit in structural equation models." In: *Multivariate applications book series. Contemporary Psychometrics: A Festschrift for Roderick P. McDonald*. Lawrence Erlbaum Associates Publishers, pp. 275–340 (cit. on p. 38).

Martin, D. B., Hanrahan, B. V., O'Neill, J., and Gupta, N. (2014). "Being a turker." In: *Computer Supported Cooperative Work, CSCW '14, Baltimore,* *MD, USA, February 15-19, 2014*. Ed. by Fussell, S. R., Lutters, W. G., Morris, M. R., and Reddy, M. ACM, pp. 224–235. ISBN: 978-1-4503-2540-0. DOI: 10.1145/2531602.2531663. URL: http://doi.acm.org/10.1145/2531602.2531663 (cit. on p. 60).

Martin, D., Carpendale, S., Gupta, N., Hoßfeld, T., Naderi, B., Redi, J., Siahaan, E., and Wechsung, I. (2017). "Understanding the crowd: Ethical and practical matters in the academic use of crowdsourcing." In: *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Springer, pp. 27–69 (cit. on p. 55).

Marvit, M. Z. (2014). *How crowdworkers became the ghosts in the digital machine*. https://www.thenation.com/article/how-crowdworkers-became-ghosts-digital-machine/. Accessed: 2019-12-20 (cit. on pp. 7, 8).

McDonald, R. P. and Ho, M.-H. R. (2002). "Principles and practice in reporting structural equation analyses." In: *Psychological Methods* 7.1, p. 64 (cit. on pp. 38, 39).

Miles, J. and Shevlin, M. (2007). "A time and a place for incremental fit indices." In: *Personality and Individual Differences* 42.5, pp. 869–874 (cit. on p. 39).

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge (cit. on p. 40).

Millsap, R. E. and Olivera-Aguilar, M. (2012). "Investigating measurement invariance using confirmatory factor analysis." In: *Handbook of Structural Equation Modeling*. Guilford Press, pp. 380–392 (cit. on p. 40).

Mimno, D. M., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). "Polylingual topic models." In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, pp. 880–889. URL: https://www.aclweb.org/anthology/D09-1092/ (cit. on p. 32).

Mitchell, T. R. and O'Reilly, C. (1983). "Managing poor performance and productivity in organizations." In: *Research in Personnel and Human Resources Management* 1, pp. 201–234 (cit. on p. 193).

Naderi, B. (2018). *Motivation of workers on microtask crowdsourcing platforms*. Springer (cit. on pp. 52, 59).

Naderi, B., Wechsung, I., Polzehl, T., and Möller, S. (2014). "Development and validation of extrinsic motivation scale for crowdsourcing micro-task platforms." In: *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia, CrowdMM '14, Orlando, Florida, USA, November 7, 2014*. Ed. by Redi, J. and Lux, M. ACM, pp. 31–36. ISBN: 978-1-4503-3128-9. DOI: 10.1145/2660114.2660122. URL: http://doi.acm.org/10.1145/2660114.2660122 (cit. on p. 58).

Ni, X., Sun, J., Hu, J., and Chen, Z. (2009). "Mining multilingual topics from wikipedia." In: *Proceedings of the 18th International Conference on World Wide Web, WWW 2009, Madrid, Spain, April 20-24, 2009*. Ed. by Quemada, J., León, G., Maarek, Y. S., and Nejdl, W. ACM, pp. 1155–1156. DOI: 10.1145/1526709.1526904. URL: https://doi.org/10.1145/1526709.1526904 (cit. on p. 32).

O'Reilly, C. A. and Chatman, J. A. (1994). "Working smarter and harder: A longitudinal study of managerial success." In: *Administrative Science Quarterly*, pp. 603–627 (cit. on p. 192).

Osborne, J. W., Costello, A. B., and Kellow, J. T. (2008). "Best practices in exploratory factor analysis." In: *Best Practices in Quantitative Methods*. Sage Thousand Oaks, CA, pp. 86–99 (cit. on p. 35).

Paolacci, G. and Chandler, J. (2014). "Inside the Turk: Understanding Mechanical Turk as a participant pool." In: *Current Directions in Psychological Science* 23.3, pp. 184–188 (cit. on p. 53).

Paolacci, G., Chandler, J., and Ipeirotis, P. G. (2010). "Running experiments on Amazon Mechanical Turk." In: *Judgment and Decision Making* 5.5, pp. 411–419 (cit. on p. 53).

Parent, G. and Eskenazi, M. (2010). "Clustering dictionary definitions using Amazon Mechanical Turk." In: *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, pp. 21–29 (cit. on p. 50).

Pavlick, E., Post, M., Irvine, A., Kachaev, D., and Callison-Burch, C. (2014). "The language demographics of Amazon Mechanical Turk."

In: *Transactions of the Association for Computational Linguistics* 2, pp. 79–92 (cit. on pp. 50, 52).

Peer, E., Brandimarte, L., Samat, S., and Acquisti, A. (2017). "Beyond the Turk: Alternative platforms for crowdsourcing behavioral research." In: *Journal of Experimental Social Psychology* 70, pp. 153–163 (cit. on pp. 52, 54).

Pesole, A., Brancati, M., Fernández-Macías, E., Biagi, F., and González Vázquez, I. (2018). *Platform workers in Europe*. Tech. rep. Luxembourg: Publications Office of the European Union (cit. on p. 10).

Pinder, C. C. (2014). *Work motivation in organizational behavior*. Psychology Press (cit. on pp. 44, 192, 193).

Pontin, J. (2007). *Artificial intelligence, with help from the humans*. `https://www.nytimes.com/2007/03/25/business/yourmoney/25Stream.html?smid=tw-share`. Accessed: 2020-02-20 (cit. on p. 5).

Porter, L. W. and Lawler, E. E. (1968). *Managerial attitudes and performance*. Homewood, IL: Irwin-Dorsey (cit. on p. 44).

Posch, L., Bleier, A., Flöck, F., and Strohmaier, M. (2018). "Characterizing the global crowd workforce: A cross-country comparison of crowdworker demographics." In: *arXiv preprint arXiv:1812.05948* (cit. on p. 17).

Posch, L., Bleier, A., Lechner, C. M., Danner, D., Flöck, F., and Strohmaier, M. (Sept. 2019). "Measuring motivations of crowdworkers: The Multidimensional Crowdworker Motivation Scale." In: *ACM Transactions on Social Computing* 2.2, 8:1–8:34. ISSN: 2469-7818. DOI: `10.1145/3335081`. URL: `https://doi.org/10.1145/3335081` (cit. on p. 19).

Posch, L., Bleier, A., Schaer, P., and Strohmaier, M. (2015). "The Polylingual Labeled Topic Model." In: *KI 2015: Advances in Artificial Intelligence*. Ed. by Hölldobler, S., Krötzsch, M., Peñaloza, R., and Rudolph, S. Vol. 9324. Lecture Notes in Computer Science. Springer, pp. 295–301. ISBN: 978-3-319-24488-4. DOI: `10.1007/978-3-319-24489-1_26`. URL: `https://doi.org/10.1007/978-3-319-24489-1_26` (cit. on pp. 14, 16).

Posch, L., Bleier, A., and Strohmaier, M. (2017). "Measuring motivations of crowdworkers: The Multidimensional Crowdworker Motivation

Scale (first version)." In: *arXiv preprint* http://arxiv.org/abs/1702.01661v1 (cit. on p. 59).

Posch, L., Schaer, P., Bleier, A., and Strohmaier, M. (2016). "A system for probabilistic linking of thesauri and classification systems." In: *KI – Künstliche Intelligenz* 30.2, pp. 193–196. DOI: 10.1007/s13218-015-0413-9. URL: https://doi.org/10.1007/s13218-015-0413-9 (cit. on pp. 16, 64, 68, 181, 185).

Potthast, M., Stein, B., Barrón-Cedeño, A., and Rosso, P. (2010). "An evaluation framework for plagiarism detection." In: *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, pp. 997–1005 (cit. on p. 50).

Prassl, J. (2018). *Humans as a service: The promise and perils of work in the gig economy*. Oxford University Press (cit. on p. 10).

Putnick, D. L. and Bornstein, M. H. (2016). "Measurement invariance conventions and reporting: The state of the art and future directions for psychological research." In: *Developmental Review* 41, pp. 71–90 (cit. on p. 40).

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora." In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, EMNLP 2009, 6-7 August 2009, Singapore, A meeting of SIGDAT, a Special Interest Group of the ACL*. ACL, pp. 248–256. URL: https://www.aclweb.org/anthology/D09-1026/ (cit. on p. 31).

Rao, L. (2009). *TC50: Crowdflower crowdsources mundane labor to the cloud*. https://techcrunch.com/2009/09/15/tc50-crowdflower-crowdsources-mundane-labor-to-the-cloud/. Accessed: 2019-12-20 (cit. on p. 6).

Raykar, V. C. and Yu, S. (2012). "Eliminating spammers and ranking annotators for crowdsourced labeling tasks." In: *Journal of Machine Learning Research* 13.Feb, pp. 491–518 (cit. on p. 48).

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. (2010). "Learning from crowds." In: *Journal of Machine Learning Research* 11.Apr, pp. 1297–1322 (cit. on p. 48).

Reinemann, C., Aalberg, T., Esser, F., Strömbäck, J., and de Vreese, C. H. (2016). "Populist political communication: Toward a model of its causes, forms, and effects." In: *Populist Political Communication in Europe*. New York, NY: Routledge, pp. 12–25 (cit. on p. 76).

Rogstadius, J., Kostakos, V., Kittur, A., Smus, B., Laredo, J., and Vukovic, M. (2011). "An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets." In: *Fifth International AAAI Conference on Weblogs and Social Media* (cit. on p. 60).

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). "The author-topic model for authors and documents." In: *Proceedings of the 20th Conference of Uncertainty in Artificial Intelligence*. AUAI Press, pp. 487–494 (cit. on p. 33).

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., and Tomlinson, B. (2010). "Who are the crowdworkers? Shifting demographics in Mechanical Turk." In: *Proceedings of the 28th International Conference on Human Factors in Computing Systems, CHI 2010, Extended Abstracts Volume, Atlanta, Georgia, USA, April 10-15, 2010*. Ed. by Mynatt, E. D., Schoner, D., Fitzpatrick, G., Hudson, S. E., Edwards, W. K., and Rodden, T. ACM, pp. 2863–2872. ISBN: 978-1-60558-930-5. DOI: 10.1145/1753846.1753873. URL: http://doi.acm.org/10.1145/1753846.1753873 (cit. on pp. 8, 9, 52).

Rusu, O., Halcu, I., Grigoriu, O., Neculoiu, G., Sandulescu, V., Marinescu, M., and Marinescu, V. (2013). "Converting unstructured and semi-structured data into knowledge." In: *2013 11th RoEduNet International Conference*. IEEE, pp. 1–4 (cit. on p. 28).

Ryan, R. M. and Deci, E. L. (2000). "Intrinsic and extrinsic motivations: Classic definitions and new directions." In: *Contemporary Educational Psychology* 25.1, pp. 54–67 (cit. on pp. 41–43).

Ryan, R. M. and Deci, E. L. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Publications (cit. on pp. 41, 42).

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., and King, J. (2006). "Reporting structural equation modeling and confirmatory factor analysis results: A review." In: *The Journal of Educational Research* 99.6, pp. 323–338 (cit. on pp. 38, 39).

Shapiro, D. N., Chandler, J., and Mueller, P. A. (2013). "Using Mechanical Turk to study clinical populations." In: *Clinical Psychological Science* 1.2, pp. 213–220 (cit. on p. 53).

Shashidhar, V., Pandey, N., and Aggarwal, V. (2015). "Automatic spontaneous speech grading: A novel feature derivation technique using the crowd." In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1085–1094 (cit. on p. 47).

Shen, W., Wang, J., and Han, J. (2014). "Entity linking with a knowledge base: Issues, techniques, and solutions." In: *IEEE Transactions on Knowledge and Data Engineering* 27.2, pp. 443–460 (cit. on p. 49).

Simpson, E. D., Venanzi, M., Reece, S., Kohli, P., Guiver, J., Roberts, S. J., and Jennings, N. R. (2015). "Language understanding in the wild: Combining crowdsourcing and machine learning." In: *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, pp. 992–1002 (cit. on pp. 48, 49).

Sint, R., Schaffert, S., Stroka, S., and Ferstl, R. (2009). "Combining unstructured, fully structured and semi-structured information in semantic wikis." In: *CEUR Workshop Proceedings*. Vol. 464, pp. 73–87 (cit. on p. 28).

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A. Y. (2008). "Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 254–263 (cit. on p. 48).

Steenkamp, J.-B. E. and Baumgartner, H. (1998). "Assessing measurement invariance in cross-national consumer research." In: *Journal of Consumer Research* 25.1, pp. 78–90 (cit. on p. 40).

Steinmetz, H. (2013). "Analyzing observed composite differences across groups." In: *Methodology* (cit. on p. 41).

Steyvers, M. and Griffiths, T. (2007). "Probabilistic topic models." In: *Latent Semantic Analysis: A Road to Meaning.* Laurence Erlbaum, pp. 424–440 (cit. on p. 29).

Stier, S., Bleier, A., Bonart, M., Mörsheim, F., Bohlouli, M., Nizhegorodov, M., Posch, L., Maier, J., Rothmund, T., and Staab, S. (2018a). *Social media monitoring for the German federal election 2017*. GESIS Data Archive, Cologne. Data file. URL: http://dx.doi.org/10.4232/1.12992 (cit. on pp. 15, 17).

Stier, S., Bleier, A., Bonart, M., Mörsheim, F., Bohlouli, M., Nizhegorodov, M., Posch, L., Maier, J., Rothmund, T., and Staab, S. (2018b). "Systematically monitoring social media: The case of the German federal election 2017." In: *arXiv preprint arXiv:1804.02888* (cit. on pp. 15, 17, 26, 51, 182).

Stier, S., Posch, L., Bleier, A., and Strohmaier, M. (2017). "When populists become popular: Comparing Facebook use by the right-wing movement Pegida and German political parties." In: *Information, Communication & Society* 20.9, pp. 1365–1388 (cit. on pp. 14, 16, 51).

Sun, J. (2005). "Assessing goodness of fit in confirmatory factor analysis." In: *Measurement and Evaluation in Counseling and Development* 37.4, pp. 240–256 (cit. on pp. 38, 39).

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). "Hierarchical Dirichlet processes." In: *Journal of the American Statistical Association* 101.476, pp. 1566–1581. DOI: 10.1198/016214506000000302. eprint: https://doi.org/10.1198/016214506000000302. URL: https://doi.org/10.1198/016214506000000302 (cit. on p. 30).

Teh, Y. W., Newman, D., and Welling, M. (2006). "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation." In: *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*. Ed. by Schölkopf, B., Platt, J. C., and Hofmann, T. MIT Press, pp. 1353–1360 (cit. on p. 31).

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* American Psychological Association (cit. on p. 34).

Thurstone, L. L. (1947). *Multiple-factor analysis; a development and expansion of the vectors of mind.* University of Chicago Press (cit. on p. 35).

Towne, W. B., Rosé, C. P., and Herbsleb, J. D. (2016). "Measuring similarity similarly: LDA and human perception." In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 8.1, p. 7 (cit. on p. 50).

Tremblay, M. A., Blanchard, C. M., Taylor, S., Pelletier, L. G., and Villeneuve, M. (2009). "Work Extrinsic and Intrinsic Motivation Scale: Its value for organizational psychology research." In: *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* 41.4, p. 213 (cit. on pp. 45, 59).

Ullman, J. B. and Bentler, P. M. (2003). "Structural equation modeling." In: *Handbook of Psychology*, pp. 607–634 (cit. on pp. 34, 37).

Vakharia, D. and Lease, M. (2015). "Beyond Mechanical Turk: An analysis of paid crowd work platforms." In: *Proceedings of the iConference* (cit. on p. 6).

Van Knippenberg, D. (2000). "Work motivation and performance: A social identity perspective." In: *Applied Psychology: An International Review* 49.3, pp. 357–371 (cit. on p. 192).

Vandenberg, R. J. and Lance, C. E. (2000). "A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research." In: *Organizational Research Methods* 3.1, pp. 4–70 (cit. on pp. 40, 41).

Venanzi, M., Guiver, J., Kazai, G., Kohli, P., and Shokouhi, M. (2014). "Community-based Bayesian aggregation models for crowdsourcing." In: *Proceedings of the 23rd International Conference on World Wide Web*. ACM, pp. 155–164 (cit. on p. 48).

Vroom, V. H. (1964). *Work and motivation*. Vol. 54. Wiley New York (cit. on p. 44).

Waas, B., Liebman, W. B., Lyubarsky, A., and Kezuka, K. (2017). *Crowdwork - A comparative law perspective*. Bund-Verlag (cit. on pp. 2, 8–10).

Wang, X. and McCallum, A. (2006). "Topics over time: A non-Markov continuous-time model of topical trends." In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, pp. 424–433 (cit. on p. 33).

Webster, J., Trevino, L. K., and Ryan, L. (1993). "The dimensionality and correlates of flow in human-computer interactions." In: *Computers in Human Behavior* 9.4, pp. 411–426 (cit. on p. 58).

Weglarz, G. (2004). "Two worlds data-unstructured and structured." In: *DM REVIEW* 14, pp. 19–23 (cit. on p. 28).

Weinberg, J. D., Freese, J., and McElhattan, D. (2014). "Comparing data characteristics and results of an online factorial survey between a population-based and a crowdsource-recruited sample." In: *Sociological Science* 1 (cit. on p. 53).

Worthington, R. L. and Whittaker, T. A. (2006). "Scale development research: A content analysis and recommendations for best practices." In: *The Counseling Psychologist* 34.6, pp. 806–838 (cit. on pp. 36, 39).

Yan, R., Gao, M., Pavlick, E., and Callison-Burch, C. (2014). "Are two heads better than one? Crowdsourced translation via a two-step collaboration of non-professional translators and editors." In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1134–1144 (cit. on p. 50).

Yuen, M.-C., King, I., and Leung, K.-S. (2011). "A survey of crowdsourcing systems." In: *2011 IEEE Third International Conference on Social Computing*. IEEE, pp. 766–773 (cit. on p. 46).

Zapilko, B., Schaible, J., Mayr, P., and Mathiak, B. (2013). "TheSoz: A SKOS representation of the thesaurus for the social sciences." In: *Semantic Web* 4.3, pp. 257–263. DOI: `10.3233/SW-2012-0081`. URL: `http://dx.doi.org/10.3233/SW-2012-0081` (cit. on p. 68).