ISSN 2510-2591

Reports of the European Society for Socially Embedded Technologies

volume 5 issue 1
2021

Proceedings of 19th European Conference on Computer-Supported Cooperative Work

Guest Editors

Antonietta Grasso, Naver Labs Europe, France
Kevin Crowston, Syracuse University, USA
Maurizio Teli, Aalborg University, Denmark
Nelson Tenório, UniCesumar, Brazil

Series Editor

Michael Koch

Impressum

The **'Reports of the European Society for Socially Embedded Technologies'** are an online report series of the European Society for Socially Embedded Technologies (EUSSET). They aim to contribute to current research discourses in the fields of 'Computer-Supported Cooperative Work', 'Human-Computer-Interaction' and 'Computers and Society'.

The 'Reports of the European Society for Socially Embedded Technologies' appear at least one time per year and are exclusively published in the Digital Library of EUSSET (https://dl.eusset.eu/). The main language of publication is English.

ISSN 2510-2591

https://www.eusset.eu/report-series/

EUSSET is an institute of Social Computing e.V., a non-profit association according to the German legal system – founded on November 13th 2012 in Bonn, Germany (Nordrhein-Westfalen Amtsgericht Bonn VR 9675).

c/o Prof. Dr. Volker Wulf
Fakultät III
Universität Siegen
57068 Siegen
E-Mail: volker.wulf@uni-siegen.de

Table of Contents

Exploratory Papers

Notes

Tejaswini Yeleswarapu, Pranav Nair, Nimmi Rangaswamy (2021): "Should We Meet IRL": Gauging Matches in Virtual Reality. In: Proceedings of the 19th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies - Exploratory Papers, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591), DOI: 10.18420/ecscw2021_ep07

"Should We Meet IRL": Gauging Matches in Virtual Reality

Tejaswini Yeleswarapu, Pranav Nair, Nimmi Rangaswamy
International Institute of Information Technology (IIIT-H), Hyderabad
Contact Author: t.yeleswarapu@research.iiit.ac.in

Abstract. Virtual Reality has evolved as a powerful, embedded and immersive technology medium to transform dating experiences. However, there is no rigorous CSCW research examining 'dating' in VR, despite social interaction being a serious topic of exploration. We aim to push the CSCW discourse on social interaction further by analyzing the dynamics of romantic reciprocality in a *fully immersive* VR application. Through a qualitative study of 30 participants in 15 pairs, we examine a customizable VR application *'RecRoom'* as a dating technology medium to analyze how dimensions of interaction - including but not limited to voice, haptics and spatiality - influence dynamics of dating experiences. We employ Tinder as a contrasting chat based medium to situate and deepen our learnings about dating in VR. Our study finds VR allowing users to efficiently and effectively 'gauge' matches resulting in well informed decisions to meet (or not) virtual partners 'IRL' or in real life than existing chat based mediums like Tinder. We believe this leads to improved experience of first dates.

Introduction

As new technologies are introduced, innovators, enthusiasts and entrepreneurs adapt them to a variety of use cases to exploit and amplify new affordances and experiences. Dating is no stranger to the above statement and dating technology has evolved to expand deeper into virtual matchmaking. With the advent of print and mass media came matchmaking newspaper advertisements; with videotaping came VHS dating; and with the web came Match.com. Now, with the mobile phone - we have Tinder. Not only has technology adapted to suit dating, dating cultures in turn adapted to the 'technology of the era' (Sales and Bishop, 2018). In India, dating is a fairly recent phenomenon gaining momentum with the rise of mobile phone adoption since the 2000s as mobile interactions afforded privacy to dating practices. Tinder brought favourable shifts to a culture obsessing over commitment in viewing temporary hookups favourably due to the non-judgemental characteristic and comfort of the online space (Newett et al., 2017). With the advent of Virtual Reality (henceforth VR) new affordances, features and public appeal are imminent in the domain of dating technologies - and VR as technology disrupting the 'right swipe' (David and Cambre, 2016). What started out, for the authors of this paper, as a study focused on VR practices for matchmaking and disrupting traditional digital dating, transformed into a research opportunity to study the specific structuring of romantic interaction in fully immersive VR platforms. The latter was carried out specifically in contrast to chat based mediums like Tinder to appraise consequences for striking romantic relationships. The authors believe this paper to be a pioneering attempt in the critical evaluation of a *fully immersive* (Castronovo et al., 2013) VR dating experience.

Numerous CSCW focused studies have examined romantic interaction in *non-immersive* virtual multiplayer games (Zytko et al., 2015; Pace et al., 2010; Zhang, 2014; Huynh et al., 2013) and intimacy in virtual worlds like 'Second Life' (Boellstorff, 2015). However, social interaction in *fully immersive* virtual environments has been investigated mainly as avatar-based systems (Bente et al., 2008; Blanchard et al., 1990; Latoschik et al., 2017; Roth et al., 2016) and social channels such as gaze behavior, non verbal behaviours and their impact on communication quality (Garau et al., 2003; Bailenson et al., 2005). While 'Second Life' (Boellstorff, 2015) provides an understanding of intimacy in *non-immersive* (Castronovo et al., 2013) collaborative social virtual environments, our study investigates added dimensions of *complete bodily immersion* driving romantic interaction in VR and its role in the evolution of dating technologies enriching interactive possibilities. The authors view VR as a stage to set the dating experience as our findings primarily revolve around interactions that are a result of introducing VR into online dating. The latter parts of the paper further build on the findings and literature to critically examine if and how VR as a rich, immersive and interactive platform supporting dating experiences, enhances and enriches the experience of 'gauging' or screening potential partners or 'matches' (David and Cambre, 2016) prior to a first date- especially in contrast to existing chat-based

multimedia platforms like Tinder that are weaker (Marcus, 2016; Daft and Wiginton, 1979) in supporting rich social interaction. It is important to note that we investigate the efficacy of VR and Tinder as standalone mediums, only focusing on interaction and not the matchmaking process.

We set up a study with 30 heterosexual participants in 15 pairs, who were made to use both Tinder and a custom VR platform from an existing social VR game called *RecRoom*. We interviewed participants on various aspects of both environments and built a comparison based on findings. Some of our findings push the boundaries of human - VR interaction by examining them in the context of dating; VR allowed for more efficient and effective gauging of matches due to:

1. More unfettered interactions by diffusing focus from the person to the environment, alleviating social tensions inherent in a dating context.

2. Engaging interpersonal and spontaneous interactions augmenting intimacy.

3. Interaction in real time, accentuating intricacies of body language and conversational nuance

4. Better 'Avenues of attraction' through bodily immersion.

5. Security of the virtual while mirroring the perspicacity of a real life date.

The above characteristics of VR aid dating partners make well informed decisions on whether or not to meet in real life, thereby improving the quality of 'first (real) dates'. However, participants reported aspects of temporality like dedicated time commitment, inability to multi task and the animated design of *RecRoom* depleting some of the experiences of immersive VR. Our paper employs an inductive approach, deriving themes from a close reading and analysis of primary data from participant interviews and serves as a preliminary step to initiate further research on the potential of fully immersive VR as a dating technology,

Related Work

Introduction to Virtual Reality

A nuanced understanding of VR systems includes an important factor known as 'immersion' that divides all modern VR systems on the basis of their ability to 'immerse' users into more real experiences. Immersion is formally defined as "the extent to which the senses are engaged by the mediated environment" and is determined by system affordances. Ivan Suthermland first introduced key concepts of immersion and sensory input and output in a simulated world - the basis of current VR research. Most VR configurations fall into three main categories with each category being ranked by its degree of immersion. *Non Immersive* VR is the simplest form of virtual reality where users interact with the environment using a conventional monitor without being immersed; *Semi Immersive* systems include

large, multiple screens or monitors that provide a medium to high level of immersion. These systems are improved versions of desktop (non immersive) VR, supporting head tracking thereby improving the feeling of 'being there'; and *Fully Immersive* systems such as head-mounted displays (HMD) or CAVE TM systems with full bodily immersion; three or four walls, a projected floor, a projected ceiling which significantly or fully cover the users' field of view (Castronovo et al., 2013). These systems are in essence, the ultimate version of VR systems enhanced by audio, visual, sensory and haptic interfaces (Mandal, 2013). 'Second Life', 'Flight Simulator' and 'RecRoom' are examples of non-immersive, semi-immersive and fully immersive VR applications respectively.

Fully immersive VR also affords 'Avatars' or digital alter egos (Latoschik et al., 2017) that may act as physical representations of users and typically mimic their expressions and body language through sensors, further differentiating fully immersive VR from chat-based mediums and non immersive VR. Non-immersive VR also affords avatars, however, without the bodily immersion and is typically navigated only through voice. Avatars play a significant role in enhancing the realism in fully immersive VR.

Social Interaction in Fully Immersive VR

The main difference separating fully immersive VR systems from other traditional digital mediums is its 'three dimensionality'- a factor that brings together fully bodily immersion, interactivity and virtual presence all under one medium. Fully immersive virtual reality, a fairly recent technology that enables full bodily (three dimensional) immersion, has already made waves in the healthcare, gaming and tourism industries and only lately has made noise in the dating industry. While there are no exclusive fully immersive VR dating apps yet, it has shown immense potential in transforming dating experiences due to providing life-like experiences, emulating FtF interactions. Just like virtual worlds offering social games, many gaming companies have launched 'social VR' apps, where players around the world meet in a three dimensional virtual space for the sole purpose of socializing and conducting leisure activities with minimal gaming. Social VR apps have gained massive popularity in the last few years for providing realistic, immersive social experiences with high customizability. As an emerging technology, rigorous research on the full scope of social interaction in fully immersive VR has only recently gained traction with little to no focus dating and romantic interaction.

Fully Immersive VR vs Chat-Based CMC: What Makes VR an Enriching Medium for Interaction

Media Richness Theory

"Media Richness" refers to the range of audio, visual, verbal, and contextual information sources (Burgoon et al., 2002). The media richness of a medium depends on its capacity to process information and varies based on information

such as “immediate feedback, range and volume of cues, channels, usage, personalization and language variety” (Lee et al., 2011; Daft and Wiginton, 1979). Ramirez et. al (Ramirez Jr and Burgoon, 2004) classified three distinct forms of “computer mediated communication” (CMC) or modalities, namely, “text-based CMC, audio-based CMC and visual-based CMC”. Lee et. al (Lee et al., 2011) further elaborates that text-based CMC has the lowest capacity to process information as it lacks “environmental, spatial, visual, auditory and other sensory information”. Audio-based CMC (audio calls or audio messages) adds “aural” information but lacks visual cues and visual-based CMC provides visual information in addition to audio sources. It is important to note that in our paper, we refer to current online dating mediums as ‘chat-based’ and not ‘text-based’ as many incorporate audio-visual features like photos, gifs, emojis, bitmojis, audio call and so on, thereby differing from traditional text-based mediums like IM. Although existing chat-based mediums have brought text, audio and visual CMCs under a single modality, they still lack in “environmental, spatial and other sensory information” (Lee et al., 2011). We believe immersive VR bridges these important gaps that chat-based dating mediums are not built to address. As (Sundar et al., 2008) argue, enriching modalities, from chat based CMC to virtual reality, offer a powerful approximation of “real, non-mediated interaction”.

Information Cues and Social Presence

As mentioned above, the capacity to process information determines a medium’s richness and therefore “information cues” play an important role. Lee et. al and Daft et. al (Lee et al., 2011; Daft and Wiginton, 1979) define cues as the “communication of information through various channels such as text (spoken or written words), verbal cues (tone of voice), or nonverbal cues (physical gestures, body language)”.The lack of cues depletes a crucial factor known as “social presence” wherein the “realness” of communication is diminished and the person communicating is reduced to a mere “object” (Short et al., 1976). In order to experience greater social presence, the media richness of a medium and communication must be close to FtF interactions. Rich interactions facilitated by “immersive” modalities make for a significantly more engaging experiences (Pedersen and Liu, 2003) as common themes of frustration with the experience of online dating on chat based mediums like Tinder and OKCupid are the lack of social presence and spontaneity in conversation (Masden and Edwards, 2015). To this end, initial impressions developed through chat-based CMCs are less evolved than FtF interactions and therefore less enriching and uni-dimensional (Lee et al., 2011; Ramirez Jr and Burgoon, 2004). We extend this argument to VR as a rich immersive medium employing body language, tonal voice modulation and expressive verbal and non verbal language as cues achieving wide range and depth of social interaction.

Non-verbal Cues and Body Language

This work builds upon and extend Jeremy Bailenson's research on the effect of non-verbal cues and body language on social interaction in immersive virtual environments to dating and romantic interaction. Bailenson suggests that non verbal cues and gestures are often correlates of specific mental states- we smile when pleased, nod when we agree and touch when we are interested in someone. Intuitively tabulating and assessing non verbal behaviour is something humans do constantly in FtF conversations. With fully immersive virtual environments, interactants can assess these behaviours with greater precision to augment normal intuitions about body language and non verbal cues occurring during social interactions (Bailenson et al., 2005). The unavailability of non-verbal cues and body language in chat based mediums diminishes the quality of interaction in a dating context as romantic interaction involves indirect cues [such as physical touch, gaze and so on] that, as mentioned above, aid in assessing or gauging crucial factors such as interest, chemistry and reciprocation [or lack thereof].

Background

The following sections offer a background on the general affordances of mobile dating apps, mechanisms of Tinder matchmaking and insights into the unique setting the study was conducted in.

Affordances of Mobile Dating

Mobile dating apps such as Tinder include "communicative affordances" (Lutz and Ranzini, 2017) that differ from traditional online dating mediums such as Match.com (Marcus, 2016). Shrock et. al (Schrock, 2015) propose four key communicative affordances - "portability", "availability", "locatability", "multimediality" - that mobile dating apps rely on. "Portability" or mobility is the key difference between mobile dating apps like Tinder and desktop-based dating apps like Match.com; Tinder can be used "on the move" in different locations in both public and private spaces while the latter can only be used in private spaces.

"Availability" of mobile media enables easy access, leading to higher frequency of usage, "Locatability" allows for easy matching, texting and meeting with users that are in the same vicinity (physical proximity) (Lutz and Ranzini, 2017; Marcus, 2016), and "Multimediality" includes three modes of communication - text, emojis, gifs, memes, audio and video calls (Lutz and Ranzini, 2017; Marcus, 2016). Tinder also relies on an additional visual affordance of photos. According to Marcus et. al (Marcus, 2016), users rely on "limited" information while swiping due to the "heavy reliance" on photos. Another important factor known as "synchronicity" or "the short amount of time in which messages are sent" (Marcus, 2016) calls for "spontaneity" and "availability" from users to make a quick decision on their potential partners' self presentation through

photos as well as their own (Lutz and Ranzini, 2017). Lutz et. al and Marcus et. al (Lutz and Ranzini, 2017; Marcus, 2016) argue that the affordance of "synchronicity" in addition to the limited information available on Tinder are constrictive, leading to issues such as "loss of interest",“ information overload", inability to gauge potential partners properly.

Matchmaking Mechanisms of Tinder

First launched in September 2012, Tinder paved the way for mobile dating apps by introducing the "swiping" motion to anonymously "like" or "dislike" potential matches. Users can "swipe right" (like) or "swipe left" (dislike) other users based on their respective locations [1].

Tinder requires a Facebook login to create a profile and automatically extracts information like photos, name, age and gender. Users can then choose to manually change the information along with writing a short "bio" or biography to introduce themselves and optionally link an Instagram or Spotify account. Factors such geographical location, mutual friends and common interests play a crucial role in adding the most compatible candidates to a list of potential matches. Users can then anonymously "swipe right" or "swipe left" on other users to indicate (dis)interest [2]. A 'match' is formed when two users 'swipe right' on one another to indicate interest (David and Cambre, 2016; LeFebvre, 2018). A mutual right swipe then results in a "match", enabling the two interested parties chat through private messaging within the app to help determine or 'gauge' if one or both partners desire further communication.

As a chat based CMC, Tinder offers various affordances in addition to texting in the form of photos, emojis, gifs and memes to further enhance interactivity and make the dating experience engaging. We chose Tinder as a dating technology medium in the study for its specific messenger-like capabilities and popularity as a dating/romance seeking technology unlike other chat based mediums like IM. (Its important to mention that we did not choose Tinder for its reputation as a 'hook-up' or casual dating app)

Dating in India

The Indian dating culture underwent a sudden and drastic change in 2011, with the launch of Tinder into the Indian market. Tinder marketed as a 'hook-up' app quickly became popular among the younger demographic, making dating and particularly hook-ups open and widespread. Apart from a few articles [3], there is no rigorous research examining dating in India. Accordingly, our references suggest that the norms of finding a partner are slowly changing; from a culture vehemently set on arranged marriages to gradually adopting and accepting 'dating' among the

[1] shorturl.at/lBCH3

[2] shorturl.at/lBCH3

[3] http://tiny.cc/ja62cz, http://tiny.cc/42kwtz

younger demographic. The shift in culture is resulting in the emergence of new practices, from serious to casual dating and even hook-ups. The agency to choose a partner is increasingly shifting from parents to the individuals themselves with the rise in popularity of online dating platforms like OkCupid and Tinder. This shift in agency is still fairly recent, with online dating being embraced the fastest by college students and young working professionals [4]. In lieu of this, we deliberately chose college students in the age group 18-23 as our target demographic.

Methods

The VR Environment

After a thorough investigation of various cross-platform VR social applications like 'vTime' and 'Facebook Spaces' and considering the potential for customization we chose *RecRoom* for the experiment. *RecRoom* is highly customizable, offers a personal lounge, and several in-built games in a VR environment. The room designed for the study was one among the default in-game templates that was customized to befit a romantic setting reminiscent of a typical bar-room scenario. We chose this specific setting based on the results (89%) of a poll sent out on our college forum, querying for a comfortable yet romantic space for a date.

As shown in Figure 1, the room had a personal lounge area furnished with comfortable couches surrounding a coffee table, a ping pong table and a dart board, a dim lit bar furnished with a bar table, stools and ample empty space and a stage with a functioning mike and speakers for karaoke. A 'choose your own game' cabinet with several games from paintball to charades cards to disk throw was also added.

Figure 1. A lounge area furnished with dartboards and a ping pong table (left) and a bar table with two bar stools (right) in *RecRoom*..

Additionally, Zytko el al.'s (Zytko et al., 2015) research on collaborative games was incorporated to enhance dating experiences in our VR space. The research

[4] http://tiny.cc/df92cz, http://tiny.cc/ja62cz

suggests that collaborative multiplayer games allow for indirect evaluation of game partners as potential romantic partners. Collaborative games or "Multiplayer Online Games" (MOGs) like World of Warcraft and MapleStory are posited as "inherently social environments" as the games allow multiple geographically separated users to interact with one another in real time. Romantic intimacy in MOGs have been investigated in several studies (Huynh et al., 2013; Pace et al., 2010; Zytko et al., 2015) that portray MOGs as "collaborative virtual environments where a player's actual self (versus their ideal self) naturally emerge through collaboration, coordination, and teamwork". Zytko et. al (Zytko et al., 2015) further elucidates that "in-game, non-competitive" collaborative activities like a picnic in a castle or building a virtual garden together create opportunities to explore "budding" romantic feelings. Accordingly, inspiration was drawn from Zytko et al.'s research to add the aforementioned collaborative activities to enhance the process of gauging dates. For this study, the VR environment was customized in a manner that afforded the ability to implement and explore the following factors:

1. **Immersion:** We study VR as twinning the intimacy of a real life date with the affordance of a chat like conversation. To do this we ensured users were made aware of interacting over a virtual medium accessible to each other, not only for verbal conversations but for physical interactions in *virtual 'real' time*.

2. **Security:** Due to the real time dimensionality of voice and avatars, disconnecting from the room was designed to be easy in case the dating experience became uncomfortable.

3. **Interactivity:** The technology of virtual reality affords a multiplicity of new and potential features in a dating experience. For example, a spatial expansion with activities, voice, haptic touch, avatar customisations, real time verbal cues, body language and haptics are new features that we decided to focus on. Our goal here was to observe how the new affordances are deployed by our participants in the context of virtual dating.

There are two main system features that work as affordances for a VR dating experience:

1. **Avatar:** The avatar in *RecRoom* is a virtual representation of the person that goes beyond the basic 'name-picture' representation. As shown in Figure 2, the avatar in *RecRoom* is simple - a long elongated torso, squarish cell-shaded palms and wrists with the ability to extend and contract the forefinger and thumb, and a fun ovalish face with eyes and a mouth that dynamically change. The facial and body expressions are a result of the participants' body language and voice. We decided to use default avatars that load up in Rec Room and minimized customizability to height, skin color, hair color and hairstyle to mimic participants' idea of their projected physical appearance.

2. **Haptic Feedback:** Haptic feedback is essentially the 'vibration' of VR controllers due to virtual touch. Vibrations are triggered in the controllers when one avatar 'touches' another avatar in the virtual space as a form of feedback, and any touch produces the same vibrations. It is also important to note that while the vibrations do not determine the area of touch for an avatar, the three dimensional nature of a fully immersive shared virtual space helps with connecting the vibrations to the intended area of touch by affording witnessing the actions of the other avatar in real time- for instance, touching an avatar's hand produces the same vibration as touching its face but witnessing the second avatar move their hand towards the face helps registering that the vibration is meant for the face.

While courting in person, touch is a very important dimension to show one's interest in the courtship ritual. VR affords the ability to physically move towards and virtually 'touch' the partner in the form of 'haptic feedback' while coming into contact.*RecRoom* also gives haptic feedback on touch and while picking up/dropping things or when inanimate objects are thrown at an avatar.

Figure 2. A female (left) and male (right) participant's avatar choosing their hairstyle and haircolour in *RecRoom*..

The Qualitative Study

A qualitative study of 30 (15 male and 15 female) heterosexual participants interacting with a fully immersive VR application *RecRoom* and Tinder was undertaken. Since there is no 'standard' way of designing dating in VR, we built and customised the virtual 'dating' environment to the study. The participants were aged between 18-23 and belonged to an academically highly ranked engineering institute in Hyderabad, India. The study was conducted strictly on heterosexual couples made up of Indian college students. With the older demographic still bearing a 'lingering mindset' of dating being taboo and homosexuality being illegal in India (at the time of the study) methodological adjustments were made by including only college students in the aforementioned age group and heterosexual couples; the skewed sex ratio of 1:7 females to males in the institute made selection of women more complicated.

We conducted the study in two phases - the first on Tinder and the second on the customized VR app *RecRoom*, followed by semi structured interviews covering

participant experiences in both mediums. Tinder was used as a base to further ground our learnings about VR. Our aim was to create and customize a VR environment with the end goal of dating (as in Tinder) exploiting the affordances of VR technology. We endeavored to echo the matchmaking mechanisms of Tinder for pairing participants while simultaneously controlling aspects of the study to suit VR.

Participant Recruitment and Pairing

A google form asking for basic details - name, age, contact number, email, branch of study, relationship status and availability - was sent out to all students in the institute to recruit participants. Due to a skewed sex ratio, we received 72 entries from men and 31 from women to participate in the study. Since the student cohort consisted of students between the age group 17-23, entries were filtered out based on age (since the age of consent is 18 in India), availability and relationship status (only single respondents seriously looking to date were recruited). The pairing for Tinder was done manually through preference forms sent out to all prospective participants who volunteered for the study. Two sets of preference forms were sent out - one for the female participants and one for the male participants.

Figure 3. An empty classroom converted into the VR room..

The forms included the name, picture and short bio of every participant along with a 'familiarity' option to filter out partners that were known or familiar to the participant. Unfamiliar partners were deliberately chosen to make the dating process as organic as possible. Due the aforementioned skewed sex ratio we gave

every woman an option of three-five men to pair from in a follow-up individualised preference form as the number of men were chosen based on their preferences and the familiarity option of the previous preference form. The men were asked for their preferences (preferences were filtered out based on the familiarity option) but the final choice to 'match' rested with the women - much like how the dating app Bumble (Bivens and Hoque, 2018) works - leaving us with a total of 15 pairs (30 participants) in the study. The final pairing was done by mapping the first preference of female participants to the first three preferences of the male participants. A pair was formed when the first preference of the female participant overlapped with either of the first three preferences of a male participant. Although a total of 25 pairs (50 participants) were initially matched through this process, 10 of 30 women dropped out due to scheduling issues, leaving us with a total of 15 pairs. The manual matchmaking process took great care to ensure every participant was matched with an unfamiliar partner.

Procedure

The participants were informed to download Tinder and 'swipe right' on the person they were matched with and asked to chat for three days, a typical duration for pairs on Tinder (Shatto, 2018). Our intention was to retain the VR pairing as the one on Tinder, while keeping this fact hidden from the participants. All pairs were instructed to not meet in person or talk over phone to ensure their voices would not be a 'give away'. We designed the study to retain the same pairs in both VR and Tinder in order to focus on exploring 'platform' specific variations in dating experiences. Participant anonymity was a critical factor in the VR study to ensure removal of bias, thereby guaranteeing a fresh start for pairs in both mediums - evaluating both medium as independent experiences.

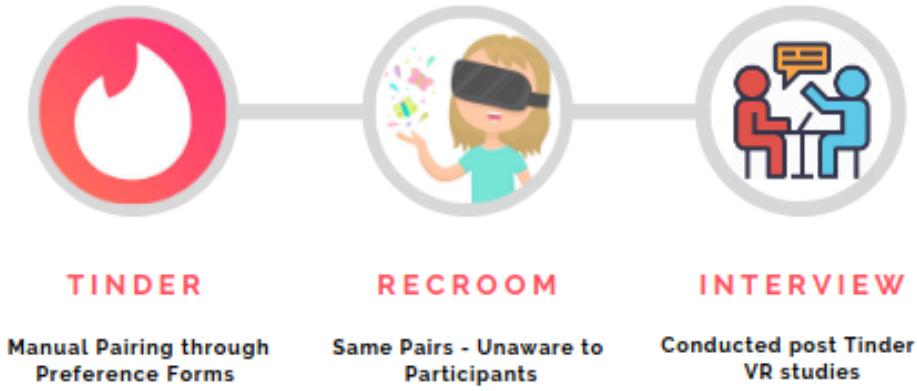


Figure 4. Chronology of the Study.

For the VR study, the pairs came in on the fourth day post the three day chat on Tinder, unaware the partner they had chatted with for the last three days will be

paired in VR too (the pairs were asked, during the interviews, if they indeed recognized their partners were identical in both mediums - only two pairs had realized that the pairing was identical). The pairs were sent to separate rooms - each of which were equipped with a VR ready PC and a VR headset - namely a HTC VIVE and an Oculus Rift. As shown in Figure 3, we used two empty classrooms, arranging sensors atop a stack of chairs and desks to capture the physical space of the room and the body of the participants. Participants were given sixty minutes [5] of in-VR time followed by a 75-90 minute detailed interview covering their dating experiences on both mediums incorporating a range of questions. To better probe into the structuring of romantic interaction affecting gauging potential, and impact of both mediums on dating experiences, we focused our interviews on the following parameters:

1. Dimensions of interactivity among paired partners - the range and depth of interaction as articulated by participants.
2. Degrees of self projection - range and richness of self articulation as expressed by participants
3. Degree of Expressiveness - how voice, haptics, touch and bodily immersion compare with emojis, gif and memes.
4. Strength of Attraction as articulated by participants.
5. Miscellaneous issues like 'security', 'control' and 'novelty' in participant experiences.

The pairs were interviewed separately by two researchers - the first and second authors of the paper. Interviews of the 30 participants and subsequent transcription were equally divided between the two researchers and the audio was recorded and transcribed with the explicit consent of the participants. Each participant received a 500 rupee (8 USD) amazon gift voucher for participation.

Data Analysis

All three authors carried out a structured, qualitative analysis to summarize and interpret interview data. Interview data was coded and analyzed using a general inductive approach (Thomas, 2006). From a careful reading of transcripts, we developed categories and and clustered excerpts together, conveying key themes from the data. Post-interviews, we transcribed the recordings and grouped qualitative opinions of every participant, based on their experiences from the study on the two modalities of dating (Tinder and VR). In the interviews, we probed the agency and intensity afforded for the five aforementioned parameters.

We then noted the qualitative opinions of participants for each of these categories under five subcategories - Supports VR, Against VR, Supports Tinder,

[5] http://tiny.cc/gy52cz

Against Tinder and Neutral and tallied the number of opinions under each of these 5 sections, for each of the five main parameters. Our qualitative interviews fleshed out the significance of the above opinions in more depth. The findings were analyzed and written from a close reading of the rich open ended interview data combined with the assessment via ratings on a 5 point Likert scale on the efficacy of gauging matches (1 being the lowest and 5 being the highest), preference (1 being strongly does not prefer and 5 being strongly prefers) and overall dating experience our participants offered about each of the two platforms.

Results

In this section, we lay out key findings from qualitative interviews about the salient features of VR influencing romantic interactions. It is important to note data pertinent to Tinder was analyzed and integrated in the findings to compare a predominantly chat based medium to VR as an immersive dating medium. Our primary focus remained VR. For quotes, we will refer to Participants as P1 to P30 while mentioning their age and gender.

The VR environment for our study offered a multi-dimensional self representation that goes beyond a picture - a full blown avatar embodying bodily presence; factors like motion controls and spatial movement adding mobility to this representation. The inclusion of voice inserted this physical representation into the social realm and the ability to communicate through various channels, including voice and touch adding an immersive layer of interaction.

Our findings demonstrate VR as a powerful preference over Tinder in terms of pure 'gauging potential' and overall 'dating experience'. A majority of participants (28/30) felt the immersive nature of VR enhanced additional opportunities to gauge matches heavily impacting the decision to schedule a first (real) date. Participants found VR to be a 'self-sufficient' and 'wholesome' medium compared to Tinder, where additional mediums like phone or video chat are often used to gauge a match before deciding to meet in real life. Of the 19 participants who wanted to schedule a date, 15 initially did not want to after the Tinder chat but changed their minds post the VR experience. Only 4 participants' decision to schedule a date did not change post the VR experience. Similarly, of the remaining 11 participants who did not want to schedule a date, 7 wanted to post the Tinder experience but changed their minds post the VR experience. Only 4 participants' decision to not schedule a date remained the same post the VR experience. As the above data explains, participants drastically modified their assessment of the same partner after the VR experience; the same partner who was deemed unworthy of a meeting after Tinder, was deemed worthy after VR and vice versa. As P15(Male, 19) succinctly puts,

> 'I do not want to meet my Tinder partner because she seemed quite distant and judgemental, which frankly was a put off and I stopped making an effort. On the other hand, I got along so well with my VR

> partner; she was intelligent, funny and our mutual love for pranks further piqued my interest in her. I'm so excited to meet for real.'

A majority of participants (26/30) did not realize (when asked) that the pairing was identical in both mediums. In extension to the previous quote, P15(Male, 19) said, 'I don't think I could've guessed my VR partner was the same as my Tinder partner - they seemed like two entirely different people!'

Participant responses are strong pointers to the constraining and abstract nature of chat to judge 'if a match on a platform is worth the effort to convert into a real life date'. As participant P27(Female, 20) put it,

> 'Meeting someone after chatting on Tinder is always dicey - you don't know if you'll hit it off for sure until you actually meet them and it's too much effort when you've matched with multiple people. VR solves this problem without having to leave one's home! It made it possible to have a life-like date which made it easier to judge if I want to take the effort to meet my match for real.'

Overall, we found participants relied and trusted the VR experience more in their decision to meet their partners due to its 'life like nature' and similarity to a real life date. The following subsections serve as further evidence to the aforementioned meta findings:

Diffusion of Focus

VR and Tinder environments bring core differences to the dating context - Tinder consists of only the screen as real estate, the chat box, the person's photos and details, and two options - check out or chat. The center of focus is the user and the pressure to keep the conversation afloat and interesting is entirely on the 'matched' (David and Cambre, 2016) pair. In VR, the environment is not restricted to a two dimensional mobile screen but extends to a three dimensional space allowing for user generated spatial quality and an environment shaped to replicate a real life scenario.

Interestingly, the spatial context in VR shifts and diffuses the focus from the person to the environment. This, we believe, afforded the alleviation of social tensions inherent in a dating environment, as the users no longer the only 'active' elements- the VR environment allowed for effective 'interaction driven props' to diffuse tensions making way for more unfettered interactions. 'Conversations (in chat) would often revolve around the participant's interests and hobbies and this unfortunately retained a sense of formalness.' - P27(Female, 20).

Participants found themselves beginning with a quick casual 'Hi' and veering headlong exploring the VR space. Indulging in virtual activities resulted not just in comfort and an air of ease but organically germinated conversations. The pairs were now provided with an immersive virtual context they mutually shared and lived in for a certain amount of time, unlike in Tinder where context had to be forged in

terms of common interests, hobbies and other related information. 21 participants found the conversation on Tinder to be 'awkwardly initiated' and the task of keeping the conversation going 'somewhat of a chore'. As participant P27(Female, 20) mentioned,

> '..as Tinder is a hit-and-miss scenario where either parties would keep introducing topics of conversation that interest them, in hopes that the other party gets hooked on to the same.'

However, 8 participants admitted that the virtual environment may have 'diffused focus on their partners than required', which led to them 'being distracted and learning less about the partner'.

Spontaneity

While chat based mediums like Tinder allow for a discontinuous, discrete messaging/ communication style, VR affords a more continuous and real time interaction. On Tinder, this discontinuity allowed participants to read messages, take time and think out replies to best impress their match and project themselves with deliberation.

> '(On Tinder)..can easily not be me, can think of replies, somebody else could be chatting, can google up facts for discussion, can even take my own sweet time..' - P9(Female, 18).

One of our participants, P15(Male, 19) compared the two dating styles to the

> 'difference between a VIVA and a written assignment.. One can rephrase in an written assignment like on Tinder, while VR was like giving a live oral examination.'

Tinder chats were 'like the take-home assignments' where the participant was allowed time to read messages, think and respond with some amount of premeditation. VR was the 'live oral exam' where participants did not get this time advantage and the interaction was 'impromptu'.

Twenty-two participants articulated the loss of spontaneity on Tinder made conversations more impersonal, nudging a waning of interest to invest in the conversation - especially when replies were temporally apart. Only seven pairs continued their Tinder conversation for the entirety of 3 days with the Tinder chats taking up not more than a total of ten minutes on any given day. On the VR platform, real time spontaneity made conversations more casual, personal and comfortable. Even when 4 out of the 15 pairs did not 'click' romantically they continued interacting in the VR environment for the entire period of 60 minutes (9 pairs had to be prodded out of the experience post the allotted time).

Immediate Feedback and Extra Linguistic Cues

Virtual Reality afforded partners the ability to give as well as receive immediate feedback on events inside of the dating experience through verbal and non verbal cues and body language expanding interactive possibilities to respond to a partner. As one participant, P4(Male, 21) put it,

> '...The combination of voice with a three dimensional representation of your partner through an avatar, touch through haptics, body language and real time interaction made the the experience intimate and wonderfully personal; conversations felt natural and life-like and doing activities like paintball and karaoke made me feel a kind of closeness and intimacy'.

18 participants voiced emotions expressed on Tinder as 'one-dimensional' and the 'props' for expression, like emojis, gifs and acronyms as 'boringly normalizing' chat in the context of texting. As P21(Male,18) said about a Tinder text,'people texting "lol" were actually not even chuckling'. In VR, participants felt their partner could not fake reactions,

> 'You can insert a facile laugh emoji on chat, but more difficult to fake a real, hearty laugh in VR' - P21(Male, 18).

While voice affords the ability to convey content, non verbal cues such as inflections and tones provide linguistic cues to elicit diverse interactions. For example, the way a word is inflected decides if it is being said with seriousness or sarcasm. However, digital forms of inflection such as emojis are not as nuanced and flexible in their use to convey emotions. Three participants alluded to 'emojis' being 'cumbersome in conveying intent and laborious in being cool and casual'. Due to the time lapse and higher probability of misunderstanding of emotions or content conveyed over chat, our participants were alert in wording messages on Tinder due to the fear of being 'immediately unmatched'. The real time conversational affordance of tenor and tone on the VR platform 'allowed for the speaker to make an immediate recovery and clear the misunderstanding before a possible unmatch.' - P4(Male, 21); 'If I use innuendo I have to put appropriate emojis to give it that tone..not required in VR' - P12(Male, 23). 29 out of 30 participants agreed that this factor of immediate feedback was a 'killer bonding' technique, as this quote highlights,

> '..Saw her shrug when I drew something disgusting, saw her shake when she laughed - could pick brilliant cues from these bodily movements. It was like I was really in the same room as her.' - P4(Male, 21)

Avenues of Attraction

Avenues of attraction are affordances to facilitate mutual attraction between the interacting partners. For Tinder - these are text, emojis, stickers, even the right-swiping experience. For VR, participants pointed out voice, chat, spatiality, mobility, avatar customisation and expressive behaviour as strong affordances.The difference in avenues of attraction in VR and Tinder stem from specific attributes of the two platforms. All 30 participants unanimously agreed VR provided more avenues for them to gauge 'partner desirability' as a date. As one of the women put it,

> 'When chatting, the person is only an idea in your head, can't gather what he's thinking' - P6(Female, 21).

Voice combined with an avatar that mimics movements and body language (smiling, talking, gestures) in real time and haptics catapulted the participants' mental conception of their match - from an abstract representation formed from a chat text to that of a 'breathing, living' person. Two of the following quotes illumine the above statement;

> 'A person's voice, her tone, her manner of speaking, her body language through her avatar, along with her intuitive response is way more attractive than an emoji- to convey the same thought' - P12(Male, 23)

> 'When the other person laughs or responds to your jokes and you can actually see them doing so, it's effective, more attractive than a chat text - P12(Male, 23).

VR also afforded an efficient way to gauge the lack of chemistry. 'Cues to connote desirability can also become cues to turn off' said, P11(Female, 19) who had noticed her match reluctant to shift from their 'joint activity' at the beginning of the 'date' or move on to some 'cosy banter'. The above 'kinds of social interactions' on the VR platform seemed to suggest to P11 '..that he was the kind to stick to his comfort zone, which could in fact be a put off'.

However, for the purpose of dating, the VR situation seemingly 'gave a lot away' as mentioned by 3 participants, P27(Female, 20) and P10(Female, 21) who did not like that 'VR was helping too much in figuring out the partner'- and preferred the control Tinder allowed to mould their 'projecting' and retained an element of m̀ystery'. One of them articulated:

> 'Certain degree of obscurity where you can't see a lot of the other person and this degree of mystery keeps you hooked. The fact you don't know how they talk, look, react..introduces a mystery keeping you attracted in a weird way' - P4(Male, 21).

Haptic feedback allowed participants to high five, bump fists, throw things at each other and especially touch - a feature adding the touch of intimacy enhancing the dating experience as an instrument of flirting.

> 'Though this isn't as real as actual physical touch, haptic touch makes flirting much more intimate and exciting, especially to subtly indicate that you're into your partner. It also made activities like paintball more fun and interactive as you actually feel the paint being thrown at you.' -P11(Female, 19)

All of the above factors afford VR an element of 'bodily immersion'. All 30 participants unanimously agreed that 'VR made them appear more attractive' because it not only changed the setting of the conversation, but also made them 'more real and fun' as the following quote illumed,

> 'On Tinder, you see my picture and my bio. In VR, you see ME' - P10(Female, 21)

Control of Experience

By control, we refer to the extent of power a user of a match making platform has over the dating experience - specifically with the ability to control conversations and the agency to be able to quit the experience at will. Tinder offered the security of 'a safe zone' - of not actually meeting a stranger in the real world and with the control to block/unmatch if the experience turned unpleasant. Participants unanimously 'upvoted' Tinder on the 'control' scale. As a participant said,

> '..On Tinder, you can immediately unmatch the person.. In VR, however, it's uneasy to remove the headset since my partner was interacting with me right there, in real time - it's more personal. If I quit my partner sees me quit..Like shutting the door on someone's face' - P12(Male, 23)

However, P18(Male, 19) provided a contrasting view,

> 'VR has the intensity of a real date and the security of virtual.'

The additional security and comfort of having a choice to leave at any stage cut both ways! A third of our participants did appreciate that VR was like a 'live stream', where something said before cannot be revisited,

> 'Chats on Tinder, as long as you haven't been unmatched, are permanent' - P5(Male, 20).

Control over ‘time commitment’ is also critical as interaction in VR required setting aside a dedicated amount of time- ‘something to be put up with’- in contrast to Tinder, where the ability to multi-task while chatting was assumed advantageous. Participant views were divided on the above; 6 did not take to the increased commitment and would rather deal with the more casual, time-discontinuous Tinder. 20 participants commented that since Tinder ‘eventually leads to a first date’ where both partners have to commit some time, they would prefer ‘cutting the redundant texting’ and go for increased time commitment in a VR environment. There were yet a few, 4 of them, who believed if they were really serious about finding a partner, they would not mind dedicating some fixed amount of time. The 6 participants who were only looking for a casual fling said, ‘they would rather not spend as much time and use Tinder instead’. The fact that Tinder afforded courtship of multiple potential dates (chat as a medium allowed users to communicate with multiple potential matches at the same time) furthered the advantage of this temporal separation in ‘chat-only’ dating technologies.

Discussion

Gauging Potential: Deciding to Meet IRL

Prior work on *non-immersive* virtual worlds like ‘Second Life’ examined the nature of intimacy and sex as ‘just another part of virtual life’. For ‘residents’ of Second life, romantic engagements or finding partners and relationships were part of a larger engagement with the game- and a vast majority of romantic relationships remained virtual, with no intention to extend them to real life relationships (Boellstorff, 2015). While Boellstroff explored romance and intimacy in a collaborative *non-immersive* virtual world limited to chat and ‘2-D avatars’ (Boellstorff, 2015), our study pushes the boundaries of his research analyzing complete bodily immersion on a VR platform with a purposive goal of extending relationships to the real world.

The goal of any dating platform, chat based or VR, is to facilitate an environment that would promote ‘gauging’ potential partners eventually leading to meeting IRL (in real life) or a real life meet. ‘Environment’ in this case speaks to the efficacy and affordances of the user interface - on chat based mediums(text, emojis, etc) and virtual space in VR (voice, avatars, haptics, etc) - augmenting the process of ‘gauging’ potential partners. Our observations illustrate the immersive and interactive nature in VR environments play a pivotal role in shaping romantic interactions that ultimately leave users with a more enriched and holistic impression of their partners compared to chat-based mediums.

We offer two interrelated ideas to further situate our observations; ‘hyper-awareness’ and ‘social waning’ to suggest a temporary release from social impulses typical of a dating context. Hyper-awareness suggests a loosening up of self-grooming and projection behaviours due to environment induced alleviation of

social pressures; Social waning suggests 'letting go' of learnt norms and defensive behaviours due to the spontaneous nature of VR combined with paired collaborative activities. The last subsection, 'Assessing Socio Romantic Behaviours', discusses the effectual assessment of socio-romantic behaviors for better of gauging matches. All three ideas aid in contextualizing and connoting the structuring of romantic interactions in VR.

Hyper-awareness and Social Waning

Hyper-awareness forms a major component in the development of romantic or sexual interactions as it engenders the need to be constantly vigilant and in control of one's appearance, words and actions in projecting the best possible version of oneself. This is present in chat based mediums like Tinder in the form of carefully curated photos and tailored conversations - an online 'performance' of self as a series of signals to convey a particular impression' (Donath, 2002). Although VR enables similar props and capabilities in the form of avatar customization, the obsessive need for impression management seemed to decrease. The real time interaction in the form of collaborative activities and games organically sparked romantic attraction between pairs by developing trust, inter-dependence, elucidating the ability to work as a team (Huynh et al., 2013; Zhang, 2014) while providing segues into the partner's personality. This shift in attention from the paired partner to the activity dissipated hyper awareness in VR as the "pressure to impress" receded, allowing for a more congenial, free-flowing interaction.

Access to the aforementioned games and activities coupled with spontaneity due to real time interaction gave rise to another key observation we call 'social waning'. The spontaneous nature of VR, with respect to conversations and participation in activities enabled participants to become less socially conscious. The multitude of activities allowed pairs to explore one another in different settings, without the hyper-awareness of a real date. The ever stubborn *"social gatekeepers of the mind"* loosened as participants reported partaking in activities showcasing facets of their personality, *"sometimes surprising themselves"*! Social waning not only enabled multi-dimensional exploration and understanding of a potential partner, but inadvertently led to the exploration and understanding of *"one's own social dimensions in a safe virtual space"*, rendering the dating experience on VR more holistic in nature. As most first dates are 'awkward' and 'unsettling' (Goodman and Churchill, 2007), the multidimensional understanding of a potential partner offers a sense of familiarity rendering the 'VR date' a valuable precursor to a 'first date'. Furthermore, the socially awkward amongst participants concurred about a VR date being more exciting than an actual date by overcoming inhibitions that accompany "the meeting of partners on a first date".

Assessing Socio-Romantic Behaviours

The addition of non verbal or extra linguistic cues and body language alongside verbal cues have proved advantageous, augmenting the quality of social interaction

in IVE's (Bailenson et al., 2005). Non verbal cues and gestures are often correlates of mental states. The intuitive tabulating and assessing of non verbal behaviour is a common human practice in FtF conversations (Bailenson et al., 2005). With fully immersive VR, our participants reported a higher degree of confidence in the assessment of the partner's socio-romantic behaviours than on Tinder. In a dating environment, the VR platform and accompanying immersive experience aided participants infer their date's social and romantic 'state of mind' and accordingly mould mutual response. The emphasis on inferring a romantic partner's engagement levels calls for an important observation in terms of 'fulfilment of expectations', as P25 (Male,18) elucidates

> 'The likelihood of hitting it off in real life, on an actual date is much higher after a VR conversation than a Tinder conversation, because in terms of personality, you'll more or less get what you expect. The same cannot be said about Tinder because chat conversations can be unreliable.'

Tinder, due to its limitations as a chat based medium, does not afford a chance to explore multiple aspects of a potential romantic partner and often leads to a commonly believed misconception that the 'match' deliberately misrepresented themselves when expectations of the person on chat don't align with the person in real life. On the other hand, the many affordances of VR, discussed in the previous section, offer an expansive experience with a potential match increasing the probability of 'expectations being met' on a first date.

As discussed above, VR structures and manipulates romantic interactions making people considerably less socially conscious, constrained, hence, withdrawn, in a first dating experience. The above coupled with insights gained about personality and non verbal behavioural traits amplifies the process of gauging matches in immersive VR. The latter seemingly and successfully augments understanding of 'how a match on VR would respond in a life-like setting and get along' leading to a well-informed decision on whether or not to take the VR date to the next level of meeting in an actual social context. We observe that immersive VR deflates the gap between chat based mediums and FtF interactions by increasing the efficiency and efficacy of gauging matches. This, we believe would lead to improved quality of subsequent first (real) dates.

Limitations

Though our work offers a unique qualitative understanding of dating in VR, we recognize limitations to our findings. First, our study was conducted in a restricted social environment limiting selection of participants to one college. Second, all our participants were heterosexual as the study was held in the metropolis of Hyderabad, South India, where dating culture is still heteronormative and restrictive with little credence given to gender fluid sexualities. Third, there is no

way gauge if our interview data would stay the same once VR has normalised and situated as 'everyday' technology. Also, the fact that VR headsets are still in between the 'enthusiast' and 'industrial' design stage (Norman, 2013), their transition into a market friendly commercially viable product may engender aspects that our study could not explore.

Study Design Challenges

We are aware the study design has implications on our data analysis and discuss four study design challenges in this section:

Controlled Study. Dating is a complex social behaviour contingent on contextual cues. Dating in controlled environment can lead to a loss of these context cues. While our study was 'controlled' with respect to time (typical 3 days on Tinder and 60 minutes in VR), none of the conversations on Tinder or interactions in VR were recorded. This was deliberately undertaken to ensure free flowing of interaction on both mediums and attempt to minimize implications of the Hawthorne Effect (Adair, 1984). Partner anonymity in the VR study was also just limited to name and identity, with only two participants reporting its imposition restrictive. Studying human behaviour in a dating context gives rise to privacy concerns, one that we circumvented, to some extent, by controlling the dating environment and sharing details of the study design with our participants. Blascovich et al. and Loomis et al.'s research on social interaction in collaborative virtual environments affirmed research advances in the understanding of the nuances and intricacies of social interaction, requiring a high level of experimental control while allowing for enhanced ecological validity (Blascovich et al., 2002; Loomis et al., 1999).

While there have been ethnographic studies to understand sex and intimacy in *non-immersive* virtual worlds such as 'Second Life' (Boellstorff, 2015), Boellstroff specifically describes his method as *virtual* ethnography - one that entailed virtually observing *online* personas or 'avatars' (Boellstorff, 2015) thereby dealing with privacy concerns to some extent. Another reason for a controlled study is the lack of prior research probing the interrelationship of technologies and dating behaviours in fully immersive VR systems. To begin with a controlled study seemed plausible in offering avenues for future research in diverse dating contexts (Castronovo et al., 2013).

Order of the Study. Tinder was experienced first, followed by VR to mimic the natural progression of romantic interaction in dating technologies- from existing chat based mediums like Tinder to fully immersive VR. Our observations from a pilot session of the study also suggested the implemented order of the study- participants better eased into the dating experience through a familiar dating medium like Tinder [similar to a real life situation] first and were then introduced

to an unfamiliar dating medium like fully immersive VR to maintain an organic flow of the dating process.

Novelty. The possibility of novelty of VR impacting our data was taken into account during the interviews asking participants if they owned/experienced VR before. Fifteen of the 30 participants reported owning VR headsets; 12 were familiar with VR while 3 had never experienced VR. Ten percent of our sample size reporting VR novelty, we deemed reasonable to assume novelty not unduly influencing data analysis. Futhermore, all 30 participants were individually given a brief demonstration of *RecRoom* along with a training session of the VR environment prior to the study to control the impact of novelty. All 30 participants reported having used Tinder prior to the study.

VR Room Design. Since there are no standard 'exclusive' VR dating apps, we had to rely on existing social VR applications for the purpose of this study. Although we chose a highly customizable application, *RecRoom*, to befit a romantic setting, a few (6/30) participants expressed the 'animated' design of the room, style of avatars and the very nature of 'Virtual Reality' made the VR date too 'gamified' and 'cartoonish', depleting the seriousness of gauging a match.

Conclusion

Fully Immersive VR has proved to be a great technological asset for many industries and is now expanding its reach to the dating and matchmaking market. As an emerging technology with the capacity to provide full bodily immersion, fully immersive VR separates itself from existing digital dating mediums, calling for an investigation on the workings of fully immersive VR in a dating context. Employing Tinder, an existing chat-based dating app as a contrasting medium, this paper explored the 'gauging' efficacy and efficiency of fully immersive VR through a controlled qualitative study.

The study offered further opportunities to deduce some of the fundamental differences structuring romantic interaction on the two mediums. Observations on the affordances of VR (especially the interactive space, feedback through verbal and non verbal cues and body language and touch through haptics) are consolidated as incrementally efficient and enriching in gauging matches, thereby improving the quality of subsequent real life dates. Dating in VR familiarized pairs with one another through a shared virtual space, paired collaborative activities in immersive VR and became a precursor to a first date. VR also afforded better avenues of expression, attraction, gauging a match more personal and intimate while offering a suitable platform for the socially anxious. The research analysis is also extended to establish the shortcomings of fully immersive VR - the dedicated time commitment, inability to multitask and animated design serving as negatives for some participants.

This paper aimed to primarily serve as a precursor to more evolved and rigorous research on romantic interaction in fully immersive VR, and encourage further discourse on a fairly novel but under explored facet of a technology capable of producing life-like experiences. Further studies can help extend this initial discourse to a more generalized understanding of socio-romantic interaction in immersive technologies.

Acknowledgments

We are grateful to IIIT Hyderabad for financial and infrastructural support; We thank Mukul Hase and Nishant Prabhu for intellectual support.

References

Adair, J. G. (1984): 'The Hawthorne effect: a reconsideration of the methodological artifact.'. *Journal of applied psychology*, vol. 69, no. 2, pp. 334.

Bailenson, J. N., A. C. Beall, J. Loomis, J. Blascovich, and M. Turk (2004): 'Transformed social interaction: Decoupling representation from behavior and form in collaborative virtual environments'. *Presence: Teleoperators & Virtual Environments*, vol. 13, no. 4, pp. 428–441.

Bailenson, J. N., A. C. Beall, J. Loomis, J. Blascovich, and M. Turk (2005): 'Transformed social interaction, augmented gaze, and social influence in immersive virtual environments'. *Human communication research*, vol. 31, no. 4, pp. 511–537.

Bente, G., S. Ruggenberg, N. C. Kramer, and F. Eschenburg (2008): 'Avatar-Mediated Networking: Increasing Social Presence and Interpersonal Trust in Net-Based Collaborations'. *Human Communication Research*, vol. 34, no. 2, pp. 287–318.

Bivens, R. and A. S. Hoque (2018): 'Programming sex, gender, and sexuality: Infrastructural failures in the" feminist" dating app Bumble'. *Canadian Journal of Communication*, vol. 43, no. 3, pp. 441–459.

Blanchard, C., S. Burgess, Y. Harvill, J. Lanier, A. Lasko, M. Oberman, and M. Teitel (1990): 'Reality built for two: a virtual reality tool'. In: *ACM SIGGRAPH Computer Graphics*, Vol. 24. pp. 35–36.

Blascovich, J., J. Loomis, A. C. Beall, K. R. Swinth, C. L. Hoyt, and J. N. Bailenson (2002): 'Immersive virtual environment technology as a methodological tool for social psychology'. *Psychological Inquiry*, vol. 13, no. 2, pp. 103–124.

Bly, S. and E. F. Churchill (1999): 'Design through matchmaking: technology in search of users'. *interactions*, vol. 6, no. 2, pp. 23–31.

Boellstorff, T. (2015): *Coming of age in Second Life: An anthropologist explores the virtually human*. Princeton University Press.

Bowers, J. M. (1991): 'The Janus Faces of Design: Some Critical Questions for CSCW'. In: J. M. Bowers and S. D. Benford (eds.): *Studies in Computer Supported Cooperative Work: Theory, Practice and Design*. Amsterdam, etc., pp. 333–350, North-Holland.

Burgoon, M., V. P. Denning, and L. Roberts (2002): 'Language expectancy theory'. *The persuasion handbook: Developments in theory and practice*, pp. 117–136.

Castronovo, F., D. Nikolic, Y. Liu, and J. Messner (2013): 'An evaluation of immersive virtual reality systems for design reviews'. In: *Proceedings of the 13th international conference on construction applications of virtual reality*, Vol. 47.

Daft, R. L. and J. C. Wiginton (1979): 'Language and organization'. *Academy of Management Review*, vol. 4, no. 2, pp. 179–191.

David, G. and C. Cambre (2016): 'Screened intimacies: Tinder and the swipe logic'. *Social media+ society*, vol. 2, no. 2, pp. 2056305116641976.

Donath, J. S. (2002): 'Identity and deception in the virtual community'. In: *Communities in cyberspace*. Routledge, pp. 37–68.

Garau, M., M. Slater, V. Vinayagamoorthy, A. Brogni, A. Steed, and M. A. Sasse (2003): 'The impact of avatar realism and eye gaze control on perceived quality of communication in a shared immersive virtual environment'. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. pp. 529–536.

Gerson, E. M. and S. L. Star (1986): 'Analyzing due process in the workplace'. *ACM Transactions on Office Information Systems*, vol. 4, no. 3, pp. 257–270.

Goodman, E. S. and E. F. Churchill (2007): 'After the match: mobility and first dates'. In: *Proceedings of the 2007 conference on Designing for User eXperiences*. p. 23.

Hancock, J. T. and P. J. Dunham (2001): 'Impression formation in computer-mediated communication revisited: An analysis of the breadth and intensity of impressions'. *Communication research*, vol. 28, no. 3, pp. 325–347.

Hancock, J. T., C. Toma, and N. Ellison (2007): 'The truth about lying in online dating profiles'. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. pp. 449–452.

Hitsch, G. J., A. Hortaçsu, and D. Ariely (2010): 'Matching and sorting in online dating'. *American Economic Review*, vol. 100, no. 1, pp. 130–63.

Huynh, K.-P., S.-W. Lim, and M. M. Skoric (2013): 'Stepping out of the magic circle: Regulation of play/life boundary in MMO-mediated romantic relationship'. *Journal of Computer-Mediated Communication*, vol. 18, no. 3, pp. 251–264.

Johansen, R. (1988): *Groupware. Computer Support for Business Teams*. New York and London: The Free Press.

Latoschik, M. E., D. Roth, D. Gall, J. Achenbach, T. Waltemate, and M. Botsch (2017): 'The effect of avatar realism in immersive social virtual realities'. In: *Proceedings of the 23rd ACM Symposium on Virtual Reality Software and Technology*. p. 39.

Lee, S., Y. Sun, and E. Thiry (2011): 'Do you believe in love at first sight: effects of media richness via modalities on viewers' overall impressions of online dating profiles'. In: *Proceedings of the 2011 iConference*. pp. 332–339.

LeFebvre, L. E. (2018): 'Swiping me off my feet: Explicating relationship initiation on Tinder'. *Journal of Social and Personal Relationships*, vol. 35, no. 9, pp. 1205–1229.

Loomis, J. M., J. J. Blascovich, and A. C. Beall (1999): 'Immersive virtual environment technology as a basic research tool in psychology'. *Behavior research methods, instruments, & computers*, vol. 31, no. 4, pp. 557–564.

Luff, P. and C. Heath (1998): 'Mobility in Collaboration'. In: *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work*. New York: ACM Press, pp. 305–314, ACM.

Lutz, C. and G. Ranzini (2017): 'Where dating meets data: Investigating social and institutional privacy concerns on Tinder'. *Social Media+ Society*, vol. 3, no. 1, pp. 2056305117697735.

Mandal, S. (2013): 'Brief introduction of virtual reality & its challenges'. *International Journal of Scientific & Engineering Research*, vol. 4, no. 4, pp. 304–309.

Marcus, S. (2016): 'Swipe to the right: Assessing self-presentation in the context of mobile dating applications'. In: *Annual Conference of the International Communication Association (ICA), Fukuoka, Japan*.

Markowitz, D. and J. Bailenson (2019): 'Virtual reality and communication'. *Human Communication Research*, vol. 34, pp. 287–318.

Masden, C. and W. K. Edwards (2015): 'Understanding the role of community in online dating'. In: *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. pp. 535–544.

McVeigh-Schultz, J., E. Márquez Segura, N. Merrill, and K. Isbister (2018): 'What's It Mean to Be Social in VR?: Mapping the Social VR Design Ecology'. In: *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*. pp. 289–294.

Newett, L., B. Churchill, and B. Robards (2017): 'Forming connections in the digital era: Tinder, a new tool in young Australian intimate life'. *Journal of Sociology*, pp. 1440783317728584.

Norman, D. (2013): *The design of everyday things: Revised and expanded edition*. Basic books.

Pace, T., S. Bardzell, and J. Bardzell (2010): 'The rogue in the lovely black dress: intimacy in world of warcraft'. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. pp. 233–242.

Pedersen, S. and M. Liu (2003): 'Teachers' beliefs about issues in the implementation of a student-centered learning environment'. *Educational Technology Research and Development*, vol. 51, no. 2, pp. 57.

Polzenhagen, F. and S. Frey (2017): 'Are marriages made in heaven? A cultural-linguistic case study on Indian-English matrimonials'. In: *Advances in cultural linguistics*. Springer, pp. 573–605.

Ramirez Jr, A. and J. Burgoon (2004): 'The effect of interactivity on initial interactions: the influence of information valence and modality and information richness on computer-mediated interaction'. *Communication Monographs*, vol. 71, no. 4, pp. 422–447.

Rivkin-Fish, M. (2005): *Sex in development: science, sexuality, and morality in global perspective*. Duke University Press.

Roth, D., K. Waldow, F. Stetter, G. Bente, M. E. Latoschik, and A. Fuhrmann (2016): 'SIAMC: a socially immersive avatar mediated communication platform'. In: *Proceedings of the 22nd ACM Conference on Virtual Reality Software and Technology*. pp. 357–358.

Sales, N. J. and J. Bishop (2018): 'Tinder and the Dawn of the Dating Apocalypse'.

Sambasivan, N., A. Batool, N. Ahmed, T. Matthews, K. Thomas, S. Gaytán, D. Nemer, E. Bursztein, E. Churchill, and S. Consolvo (eds.) (2019): '"They Don't Leave Us Alone Anywhere We Go": Gender and Digital Abuse in South Asia'.

Schmidt, K. and L. Bannon (1992): 'Taking CSCW seriously: Supporting articulation work'. *Computer Supported Cooperative Work (CSCW)*, vol. 1, no. 1, pp. 7–40.

Schrock, A. R. (2015): 'Communicative affordances of mobile media: Portability, availability, locatability, and multimediality'. *International Journal of Communication*, vol. 9, pp. 18.

Schuemie, M. J., P. Van Der Straaten, M. Krijn, and C. A. Van Der Mast (2001): 'Research on presence in virtual reality: A survey'. *CyberPsychology & Behavior*, vol. 4, no. 2, pp. 183–201.

Sellen, A. J. and R. H. Harper (2003): *The myth of the paperless office*. MIT press.

Seth, N. (2011): 'Online matrimonial sites and the transformation of arranged marriage in India'. In: *Virtual Communities: Concepts, Methodologies, Tools and Applications*. IGI Global, pp. 951–974.

Sharma, V., B. Nardi, J. Norton, and A. Tsaasan (2019): 'Towards Safe Spaces Online: A Study of Indian Matrimonial Websites'. In: *IFIP Conference on Human-Computer Interaction*. pp. 43–66.

Shatto, R. (2018): 'Here's How Long You Should Text Before Having A First Date, According To Experts'.

Short, J., E. Williams, and B. Christie (1976): *The Social Psychology of Telecommunications*. John Wiley and Sons Ltd.

Slater, M. and S. Wilbur (1997): 'A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments'. *Presence: Teleoperators & Virtual Environments*, vol. 6, no. 6, pp. 603–616.

Smith, H. J. and M. Neff (2018): 'Communication Behavior in Embodied Virtual Reality'. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. p. 289.

Sundar, S. S. (2008): 'The MAIN model: A heuristic approach to understanding technology effects on credibility'. *Digital media, youth, and credibility*, vol. 73100.

Sundar, S. S., A. Oeldorf-Hirsch, and A. Garga (2008): 'A cognitive-heuristics approach to understanding presence in virtual environments'. In: *PRESENCE 2008: Proceedings of the 11th Annual International Workshop on Presence*. pp. 219–228.

Thomas, D. R. (2006): 'A general inductive approach for analyzing qualitative evaluation data'. *American journal of evaluation*, vol. 27, no. 2, pp. 237–246.

Titzmann, F.-M. (2013): 'Changing patterns of matchmaking: The Indian online matrimonial market'. *Asian Journal of Women's Studies*, vol. 19, no. 4, pp. 64–94.

Toma, C. L. (2010): 'Perceptions of trustworthiness online: the role of visual and textual information'. In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. pp. 13–22.

Tu, K., B. Ribeiro, D. Jensen, D. Towsley, B. Liu, H. Jiang, and X. Wang (2014): 'Online Dating Recommendations: Matching Markets and Learning Preferences'. In: *Proceedings of the 23rd International Conference on World Wide Web*. New York, NY, USA, p. 787–792, Association for Computing Machinery.

WALTHER, J. (2001): 'Is a Picture Worth a Thousand Words? : Photographic Images in Long-Term and Short-Term Computer-Mediated Communication'. *Communication Research*, vol. 28, no. 1, pp. 105–134.

Walther, J. B. (1993): 'Impression development in computer-mediated interaction'. *Western Journal of Communication (includes Communication Reports)*, vol. 57, no. 4, pp. 381–398.

Williams, E. (1977): 'Experimental comparisons of face-to-face and mediated communication: A review.'. *Psychological Bulletin*, vol. 84, no. 5, pp. 963–976.

Xia, P., B. Liu, Y. Sun, and C. Chen (2015): 'Reciprocal Recommendation System for Online Dating'. In: *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*. New York, NY, USA, p. 234–241, Association for Computing Machinery.

Zhang, G. (2014): 'Can you marry me?: Conceptualizing in-game marriage as intimacy-mediated collaboration'. In: *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*. pp. 273–276.

Zytko, D., G. Freeman, S. A. Grandhi, S. C. Herring, and Q. G. Jones (2015): 'Enhancing evaluation of potential dates online through paired collaborative activities'. In: *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. pp. 1849–1859.

Zytko, D., S. A. Grandhi, and Q. Jones (2014): 'Impression Management Struggles in Online Dating'. In: *Proceedings of the 18th International Conference on Supporting Group Work*. New York, NY, USA, p. 53–62, Association for Computing Machinery.

Daniel Klug, Elke Schlote (2021): Designing a Web Application for Simple and Collaborative Video Annotation That Meets Teaching Routines and Educational Requirements. In: Proceedings of the 19th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591), DOI: 10.18420/ecscw2021_ep15

Designing a Web Application for Simple and Collaborative Video Annotation That Meets Teaching Routines and Educational Requirements

Daniel Klug[1], Elke Schlote[2]
[1] Carnegie Mellon University, *dklug@cs.cmu.edu*
[2] University of Basel, *elke.schlote@unibas.ch*

Abstract. Video annotation and analysis is an important activity for teaching with and about audiovisual media artifacts because it helps students to learn how to identify textual and formal connections in media products. But school teachers lack adequate tools for video annotation and analysis in media education that are easy-to-use, integrate into established teaching organization, and support quick collaborative work. To address these challenges, we followed a design-based research approach and conducted qualitative interviews with teachers to develop TRAVIS GO, a web application for simple and collaborative video annotation. TRAVIS GO allows for quick and easy use within established teaching settings. The web application provides basic analytical features in an adaptable work space. Key didactic features include tagging and commenting on posts, sharing and exporting projects, and working in live collaboration. Teachers can create assignments according to grade level, learning subject, and class size. Our work contributes further insights for the CSCW community about how to implement user demands into developing educational tools.

1 Introduction

Audiovisual media artifacts, such as music videos or YouTube clips, are an essential part of teenagers' and younger adults' everyday life and their primary means to develop digital media literacy (Lange, 2016). In line with the dramatic increase in media consumption in recent decades by kids and teenagers (McNally and Harrington, 2017), school curricula more and more require didactic approaches to teach the analysis of audiovisual media artifacts (BMBWF, 2019; Erziehungsdepartement Basel-Stadt, 2013; Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2016).

In this context, teachers are ultimately the decision makers when it comes to choosing educational tools and how to integrate them into existing teaching practices. Working with digital educational tools means teachers need to provide digital media content, design innovative learning environments, and develop a collaborative environment to engage with students in an effort to improve peer learning and teacher-student-interaction. While audiovisual media artifacts are popular learning resources in *Language* classes, *Music*, or *Arts*, teachers lack adequate digital educational tools to teach video analysis and annotation according to these new curricula standards. Teachers commonly use apps like GarageBand or iMovie, which are not primarily designed for educational purposes. Meanwhile, studies find that teachers are reluctant to use advanced analytical tools (Chien et al., 2014; Sang et al., 2011) and do not always see educational benefit in using apps recommended by educational institutions (Al-Zaidiyeen et al., 2010). In addition, Holstein et al. (2017) show teachers stop using digital tools if they cannot adapt to changes in curricula, hamper tracking student performance and following their learning process, and allow students to trick the system rather than tackling the tasks.

The main observation is that many available tools are too complex or too limited. They are either not primarily designed for use in *Language*, *Arts*, or *Music* classes, or exclusively designed for specific subjects or learning activities. However, a more crucial problem is, that teachers are generally reluctant to use digital tools because analyzing audiovisual media artifacts is a rather minor aspect in most subjects and there is only little guideline for how to adapt digital tools into established teaching routines. To ensure better accessibility and usability for educators, educational tools for video annotation need to provide the right set of features but foremost tools need to mind the didactic and methodological context in which they should be used in the first place. For example, teaching time is scarce and not intended for lengthy introductions and explanations of tools; film and video analysis in secondary education draws on a well-defined and unchangeable set of analytical categories (Bordwell, 2004) but requires fewer advanced features than in the academe; teachers and students alike need a clear motivation and benefit to work with a tool work should be quick and easy to set up with a tool. Video annotation tools need to adapt to the didactic context rather than creating new ones. These tools should not require additional effort from teachers but present easy-to-use digital options for subject-specific activities. This results in a strong need for adequate educational video annotation tools.

This paper presents the development and design of TRAVIS GO, a web application for video analysis and annotation (http://travis-go.org/en) that meets (a) the educational and didactic requirements of curricula, and (b) the needs and demands of teachers for easy integration into teaching contexts (Schlote and Klug, 2020). Our work is guided by the following research questions:

- **RQ1:** *What are didactic requirements for teaching about audiovisual media artifacts?*
- **RQ2:** *What are features teachers need in digital educational tools to teach video annotation and analysis?*
- **RQ3:** *How can a digital educational tool for video annotation and analysis easily support collaboration and teacher-student interaction?*

To answer these research questions, we followed a design-based research (DBR) approach (Design-Based Research Collective, 2003). *First*, we researched educational needs and demands for video annotation tools by analyzing curricula, *second*, we conducted expert interviews with teachers. *Third*, we developed TRAVIS GO following an iterative process of designing and redesigning the web application.

2 Collaboration and Audiovisual Media Analysis and Annotation

Shorter audiovisual media formats, such as TV series, or music videos, are largely popularized by YouTube, TikTok, or Instagram. They present "video-mediated lifestyles" (Lange, 2016) for kids and teenagers whose digital media consumption (Frees et al., 2019; McNally and Harrington, 2017) and use of digital devices continues to steadily increase (Anderson and Jiang, 2018; Rideout and Robb, 2019). Studies demonstrate that students of all ages are highly familiar and socialized with various audiovisual media content (Medienpädagogischer Forschungsverbund Südwest, 2018; Suter et al., 2018). This further illustrates the need for teachers to be able to design didactic material that includes audiovisual media artifacts related to students' media life-worlds. To achieve this, they need adequate digital tools to realize cooperative learning in groups, to support individual work strategies, and to provide direct feedback (Bundesamt für Sozialversicherung, 2019).

Asynchronous collaboration in online contexts (Cadiz et al., 2000; Dorn et al., 2015; Weng and Gennari, 2004) as well as annotation practices to collaboratively work on (Bargeron et al., 2002; Diamant et al., 2008) and with media artifacts (Crabtree et al., 2004; Hartmann et al., 2010) are long-established research areas within CSCW. More recent work discusses and evaluates collaborative tools to improve distributed teamwork and group discussions, for example, through creating visual cues (Shi et al., 2017), visualizing group

dynamics (Lim and Chiu, 2015), or sharing multimedia content in and after distributed team meetings (Marlow et al., 2016).

In educational contexts like computer-supported collaborative learning (CSCL), collaborative tools are designed, for example, to enable multimodal feedback (Yoon et al., 2016), improve students' skills to work in online groups (Ahuja et al., 2019), or to analyze the effects of collaborating in Google Docs on students' synchronous writing practices (Yim et al., 2017).

Previous work on collaboration and human-computer interaction presents various tools for annotating and analyzing media content (Burr, 2006; Cunha et al., 2013; Hosack, 2010; Liu et al., 2019) mostly designed for general (non-student) audiences (Hamilton et al., 2018; Nguyen et al., 2013). For example, tools to visualize Twitter claims (Pollalis et al., 2018), to analyze online information quality (Diakopoulos et al., 2009), to create dynamic annotations in web page text content (Hong and Chi, 2009), to automate video annotation (Wang et al., 2018), or to create multimodal annotations to improve creative processes in dance rehearsals (El Raheb et al., 2018; Singh et al., 2011).

From a learning science perspective, papers address how students learn with videos but not how they learn about videos. Papers, for example, analyze types of engagement with video in active learning (Dodson et al., 2018), peaks of activity in video learning (Kim et al., 2014), or how in-video prompting helps to prevent student disengagement in video learning (Shin et al., 2018). Numerous research focuses on collaborative aspects, such as designing online learning platforms (Alper et al., 2017) or tools that allow kids to create audiovisual projects on mobile devices (Hickey, 2019). Many tools are designed for collaborative work; however, not primarily in educational contexts. For example, tools are designed to collaboratively analyze rather specific audiovisual media artifacts, such as TV debates (Carneiro et al., 2019), or to engage in specific analytical activities, such as concept mapping in video learning (Liu et al., 2018), or collaborative video editing (Merz et al., 2018). The few collaborative educational tools mostly aim at peer learning through simple but asynchronous video annotation (Baecker et al., 2007; Singh et al., 2016) or at collaborative collection, creation, and assemblage of audiovisual media (Hamilton et al., 2018; Heimonen et al., 2013). A notable exception is the work by Chen, Freeman, and Balakrishnan (2019) who developed and evaluated a live streaming interaction tool for language teachers. Their study shows that adding video to language learning improves feedback and increases student engagement.

3 Methodology and Research Process

The development of TRAVIS GO was guided by a design-based research (DBR) approach (Design-Based Research Collective, 2003; Wang and Hannafin, 2005). This is a common method in learning sciences to study complex systems and to

generate solutions to a problem that is then tested and evaluated in practice (Barab and Squire, 2004; Cobb et al., 2003). In our case, we identified the lack of adequate educational tools that also mind the didactic and methodological context as the problem, the design process of TRAVIS GO illustrates the implications how this problem was solved.

The development of TRAVIS GO primarily focused on teachers' needs and demands for an adequate analytical tool as they are the more critical population when it comes to adapting to and introducing new tools into teaching. Teachers are generally less open to use digital tools or web apps because audiovisual media artifacts are not a main subject in most school subjects but an additional perspective, for example, when comparing a book with a film in *Language* classes. Any new educational tool requires teachers to invest prior time and effort to familiarize with its features, to design didactic use cases that are adequate for a subject, grade, and curricula demands, and to make sure they understand a tool in a way they can teach and support students in working with it. Yet, there are little to no trainings offered to help teachers, and curricula provide no guidelines of how to adapt new tools into existing teaching practices. This leads to reluctance towards new tools and rather results in teachers using apps they know but are inapt. Teachers as well worry they could lose their authority as educator if they are not able to explain and demonstrate digital educational tools to students. In contrast, students are digital natives and more skilled in using and understanding digital tools, platforms, and digital media than teachers because of their high everyday use of, for example, social media and smart digital devices. Therefore, we define teachers as the main group to target when developing a web app for annotating and analyzing audiovisual media artifacts in educational context, especially because teachers rather than students are the gatekeepers who decide if a new tool will be used in class. It is above all important to make sure digital educational tools are actually used to teach basics of how to analyze and annotate audiovisual media artifacts. This means providing tools that appeal to teachers and fit into teaching routines.

We first researched *educational needs and demands* for video annotation tools by analyzing curricula for the above-mentioned subjects. In the development of TRAVIS GO, we used the school subjects of *Languages*, *Music*, *Arts*, and *History* as examples because these commonly involve teaching with and about audiovisual media. Second, we conducted *semi-structured expert interviews* with 31 teachers. Curricula define learning goals, interactions, and learning content for a school subject across grades. Teachers are experts who can provide first-hand insights into didactic practices and working with digital tools. Figure 1 describes the number of each type of data we collected by subject, state, and grade.

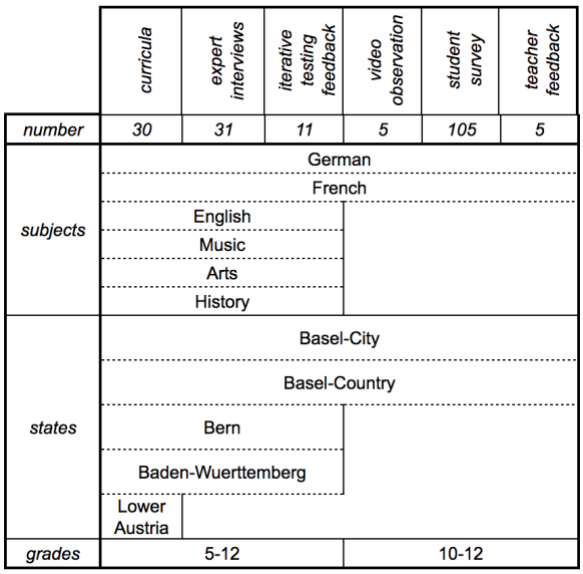


Figure 1. This figure chronologically (columns from left to right) shows the type and number of data we collected for each school subject, in each state, and for each grade (e.g. we reviewed 30 curricula in all subjects, for all states for grade 5-12.

We used qualitative document analysis (Flick, 2018) to extract all passages from the 30 curricula documents that define how audiovisual media artifacts should be integrated in *Languages*, *Music*, *Arts*, and *History* on different grades to identify key artifacts and key activities in each subject to be covered by the analytical features of the planned web application.

The 31 semi-structured expert interviews[1] (22 male, 9 female) (Bogner, Littig and Menz, 2009) provided rich information on processes and linked contexts (Meuser and Nagel, 2009), that is, needs and demand for collaborative video annotation tools. The interview questions covered motivations to use audiovisual media artifacts in teaching, application of educational standards to include audiovisual media artifacts into teaching, access to and use of technical equipment, current use of didactic audiovisual material, and expectations for a video annotation tool. In the qualitative thematic content analysis of the interview data (Flick, 2018), two experienced researchers generated initial codes through inductive coding for one interview with the software MAXQDA, compared codes, and agreed on a coding scheme. Then, one researcher proceeded to code

1 All interviews were recorded by permission, anonymized, and transcribed. All interviews were conducted in German, all quotes have been translated into analogous English.

the interview data by following the coding scheme. This process is an established way in the CSCW community to ensure validity of qualitative data through a high level of agreement (McDonald et al., 2019). The codes were then structured into nine categories: (1) *personal motivation to use audiovisual media in teaching*, (2) *institutional obligation to use audiovisual media in teaching*, (3) *technical and infrastructure aspects of teaching*, (4) *implemented didactic use of audiovisual media artifacts*, (5) *desired didactic use of audiovisual media artifacts*, (6) *implemented forms of work in teaching*, (7) *desired forms of work in teaching*, (8) *demands for using video annotation tools*, and (9) *critical assessment of using video annotation tools*. These categories served as basis for identifying key challenges in the development of TRAVIS GO (see 4) and informed the design of the web app.

According to DBR, in the process of designing and redesigning the first version of TRAVIS GO, we performed iterative tests with 11 teachers and collected their feedback through email and in conversations to evaluate the didactic and collaborative value of the web application. Because teachers only addressed minor aspects in mostly positive feedback we did not perform interventions but refined the first app version for the second and final version of TRAVIS GO.

4 Challenges to Address in the Development of TRAVIS GO

Analyzing curricula and expert interviews with teachers resulted in key challenges that were addressed in the development of TRAVIS GO in order for the web application to adequately meet user needs. We are foremost considering needs and demands of teachers as the crucial group of users. They are the gatekeepers who decide which tools they feel comfortable with to include into their teaching routines, therefore the premise is to develop a tool that improves accessibility and usability for teachers.

4.1 Curricula recommendations for integrating audiovisual media artifacts

The curricula analysis shows a general direction towards active and conscious ways to include various media forms and genres. Sampled curricula define similar standards for teaching analysis and interpretation of audiovisual media artifacts (see Figure 2), for example: analyzing films and commercials in relation to students' life world (*Arts*) (Erziehungsdepartement Basel-Stadt, 2013), using digital tools to describe film music (*Music*) (Ministerium für Kultus, Jugend und Sport Baden-Württemberg, 2016), interpreting film adaptations regarding cultural

and historical contexts (*Languages*) (Erziehungsdirektion des Kantons Bern, 2017), or discuss the manipulability of language for political propaganda purposes in historical documentaries (*History*) (Erziehungsdepartement Basel-Stadt, 2013). Our interviews show teachers generally agree with curricula guidelines and requirements (*desired forms of work in teaching*) and are eager to follow curricula as institutional precepts (*institutional obligation to use audiovisual media in teaching*). However, no curriculum explicitly names digital tools for analyzing or annotating audiovisual media artifacts.

4.2 Motivation of teachers to engage with audiovisual media artifacts

German studies show teachers rarely use digital media and apps in teaching (18%), although they feel it would increase students' motivation to learn (88%) (Rohleder, 2019). Austrian studies find only 33% of media literacy classes actually discuss or interpret (audio-)visual media artifacts (Bundes Jugend Vertretung, 2017). In the expert interviews, teachers confirmed that students appreciate activities different from regular lessons and working with tools and platforms, like Moodle: *"For me, these are valuable tools, easy to learn and they are also motivating for students"* (English teacher, Basel-City). Teachers also said showing films or videos loosens the educational setting and improves teacher-student-interaction. However, we found teachers are primarily motivated by personal interest in, e.g. films, to further engage with audiovisual media artifacts in teaching (*personal motivation to use audiovisual media in teaching*) rather than by didactic examples, educational benefits, or suggested use cases provided by curricula (*desired didactic use of audiovisual media artifacts*).

Subject	Artifact	Activity
Arts	Films Commercials	reflect describe analyze interpret
Music	Sound Phenomena Film Music Visual Art	identify describe analyze
Languages	Films Film Adaptations Music Videos	analyze interpret compare understand
History	Historical Films Documentaries	discuss compare interpret

Figure 2. Overview of the media artifacts and the activities students should learn in each of the subjects according to curricula we analyzed.

4.3 Common ways of teaching with and about audiovisual media artifacts

The analysis of educational needs and demands for video annotation tools revealed that teachers in *Languages*, *Music*, *Arts*, and *History* mainly use audiovisual media artifacts to illustrate facts and contexts. Expert interviews showed teachers most commonly illustrate other learning resources with additional audiovisual media artifacts or guide students in producing videos in various contexts. However, both activities do not teach students, for example, to understand audiovisual coherences (*implemented didactic use of audiovisual media artifacts*). Analyzing audiovisual media artifacts is subject-specific, for example, comparing filmic presentation of characters to literary sources in Language classes: *"We look at the film and ask: What is the difference to the book? But also, what role does music play because you don't have that in the book?"* (English teacher, Basel-City). But teaching analysis is limited to teachers' knowledge of film analysis methodology and programs or apps they feel comfortable with, such as GarageBand, which often were designed for different purposes. Annotating films, videos etc. is the easiest for teachers and students to approach further analysis of audiovisual media artifacts, yet it is the least common. If used in teaching, associative annotation aims at teaching students to access their cultural memory through audiovisual media content: *"In my lessons I often work associatively and try to find out how much their* [the students'] *memory of associations is already filled"* (Art teacher, Basel-Country). The interviews revealed that work scenarios (*implemented forms of work in teaching*) are rather focused on teaching with than about audiovisual media artifacts.

4.4 Organizational and infrastructure aspects of teaching with audiovisual media artifacts

In interview teachers explained they prefer easy-to-use digital tools to avoid a disproportionate amount of time needed to adapt existing learning material versus time assigned to each subject per week: *"In the end I have very, very little time for my lessons and foremost I need to look at what's the outcome"* (History teacher, Basel-City). On the bright side, in contrast to German studies (Rohleder, 2019), we found that all our test schools in Switzerland had excellent technical infrastructure (fast wireless internet, computer rooms or mobile laptops, iPads, school server) to teach video analysis and annotation (*technical and infrastructure aspects of teaching*); however, teachers were missing adequate digital tools that meet their didactic needs.

4.5 Recommended educational tools for video analysis and annotation

A brief review reveals commonly-used tools and tools recommended by educational institutions (International Society for Technology in Education, 2017; Vega and Robb, 2019) are often not primarily designed for educational use and therefore limited in teaching students how to analyze and interpret media artifacts Recommended tools are often scientific and too complex (e.g., ELAN, VideoANT, Videonot.es), educational versions of media production or editing programs (e.g., Final Cut, Audacity, Loopmash), gamified programs (e.g., ArKaos, jam2jam, Songs2see), or designed for only specific school subjects and learning activities (e.g., Better Ears, Lilypond, Sibelius) (Klug and Schlote, 2018). In addition, our interviews show that teachers do not necessarily always see the use or benefit in officially recommended tools and apps (*critical assessment of using video annotation tools*) though they generally see a need to include such tools into teaching if they better match their needs and demands (*demands for using video annotation tools*).

Overall, these key challenges demonstrate that the bigger need is to develop educational tools that appeal to teachers by fitting into institutionalized routines and not by providing more complex features or approaches or new educational designs. In order to improve the accessibility and usability of educational tools, the development needs to address these educational challenges on organizational and institutional levels rather than on analytical or methodological levels.

5 Development and Design of TRAVIS GO

Based on the challenges identified in the analysis of curricula and interview data, we developed and designed the interface and features of TRAVIS GO to meet educational needs and demands as well as teaching routines (Schlote, Klug, and Neumann-Braun, 2020). The goal was to design a video annotation tool that is easy to understand, and allows to be quickly applied within established didactic routines and technical infrastructures. This is mostly based on the fact that teachers have only limited time per lesson and rely on well-practiced teaching routines. Therefore, accessing the web application should not lead to technical difficulties or take away valuable learning time. The web application should be free to use without time or location restrictions and be compatible with all digital devices, operating systems and browsers.

5.1 User Access

In the interviews, teachers expressed the need for a digital educational tool that reduces effort in preparing and teaching lessons with audiovisual media artifacts

and does not cause additional technical or time effort: *"To be honest, set up and onboarding should take max five minutes"* (History teacher, Basel-City). Therefore, we designed TRAVIS GO as a web application that is accessible free of charge and without restrictions such as time or location. It is compatible with all digital devices, operating systems and browsers. Users only need to choose a temporal user name but do not need to register or login to use TRAVIS GO. In this way, the app reduces user and data management effort for the benefit of increasing openness, usability, inclusiveness, and fast onboarding to match organizational preconditions and didactic needs in school contexts. TRAVIS GO is available in German, English, French, and Italian. These and the following (5.1.1, 5.1.2) design implications help to solve the organizational challenges that teachers generally face in accessing and integrating digital educational tools into teaching routines.

5.1.1 Privacy and Data Storage

TRAVIS GO does not store any data and requires no registration or login but only a one-time user name. This absolves data security and privacy concerns that are exceptionally crucial for developing browser-based educational web applications in school contexts (Kumar et al., 2019), for example, when students and teachers use cloud services, such as Google Drive (Arpaci et al., 2015). Moreover, because TRAVIS GO users may be minors waiving data collection also avoids potential privacy issues in favor of inclusiveness.

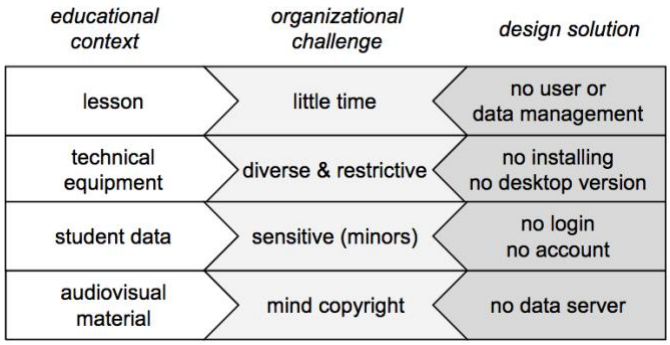


Figure 3. TRAVIS GO provides design solutions to organizational challenges teachers face when using video annotation tools in educational context.

5.1.2 Copyright and Video

TRAVIS GO also does not store any media data but only relays media through the browser, evading any possible copyright issues concerning the use of films, YouTube videos etc. as online teaching media resources. Figure 3 shows how

these organizational challenges for each educational context were solved in the design process of TRAVIS GO.

5.2 Project Management

In interviews, teachers unanimously expressed demand for an easy- and intuitive-to use tool with little to no learning curve. They voiced concern about losing authority if they are not able to quickly explain how to use the tool and its features and provide technical support to students. As a solution, to start a project in TRAVIS GO, users simply post the URL of the video or audio source (see Figure 4). To open a project, users drag-and-drop or choose the saved text file (see 5.5) from the hard drive. If users want to join a collaboration, they paste the collaboration code instead of a link (see 5.4).

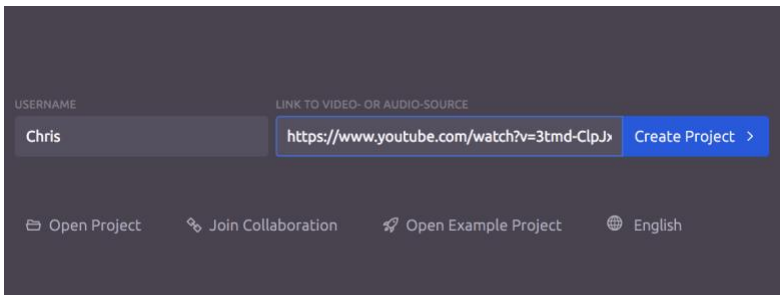


Figure 4. To start a new project in TRAVIS GO, users only need to paste a video URL (e.g. YouTube, Dropbox) on the app start page and give a temporary user name.

5.3 Work Space and Features

The interviews showed that teachers above all value accessibility, usability, and open collaborative work environments in order to appeal to students' media skills: *"In my classes we developed a very open exchange, a structure where as a teacher I'm almost dropping out. I find that very exciting"* (Arts teacher, Basel-Country). Within an openly structured work space, teachers demanded simplified analytical features compared to tools they currently used. Many teachers lack educational knowledge to teach film analysis *("The problem is missing education and knowledge in film science"*, German teacher, Basel-Country) or focus on filmic aspects, such as montage or camera movements, but not music and sound *("Music is only analyzed in a reduced way, there's no time or expertise for it"*, Art teacher, Bern). Figure 5 shows how the demand conducted in interviews played into the design of the TRAVIS GO work space and features. The *video player* (1) allows to play and navigate the selected video or audio source. In the *text editor* (2), users create annotation posts by setting a start and end point for a video sequence, and assigning an analytical category (picture, audio, text, meta)

to the sequence. Each post shows up chronologically in the *annotation feed* (5) giving the drafted text and the timestamp, category, and username in a specific syntax ([00:12:34 PICTURE @username]). Projects can be given a *title and a description* (4), they can be *searched by text* (6), *shared as text file* (7), or in a *live collaboration* through a unique collaboration ID (8) (see 5.4). Projects can be *exported and saved* as Word or text document (9) (see 5.5). The *user list* (3) shows who is currently active (in green) in or contributed (in grey) to the project. Analytical features in TRAVIS GO are universal, they derive from film analysis (Bordwell, 2004) and allow to describe and analyze characteristics and effects of audiovisual media artifacts. The feature design enables subject-specific and case-sensitive adaptation without being technically deterrent for users. This methodological design allows for advancing common ways of teaching by providing a set of general analytical features that match teachers' desired didactic use of audiovisual media artifacts.

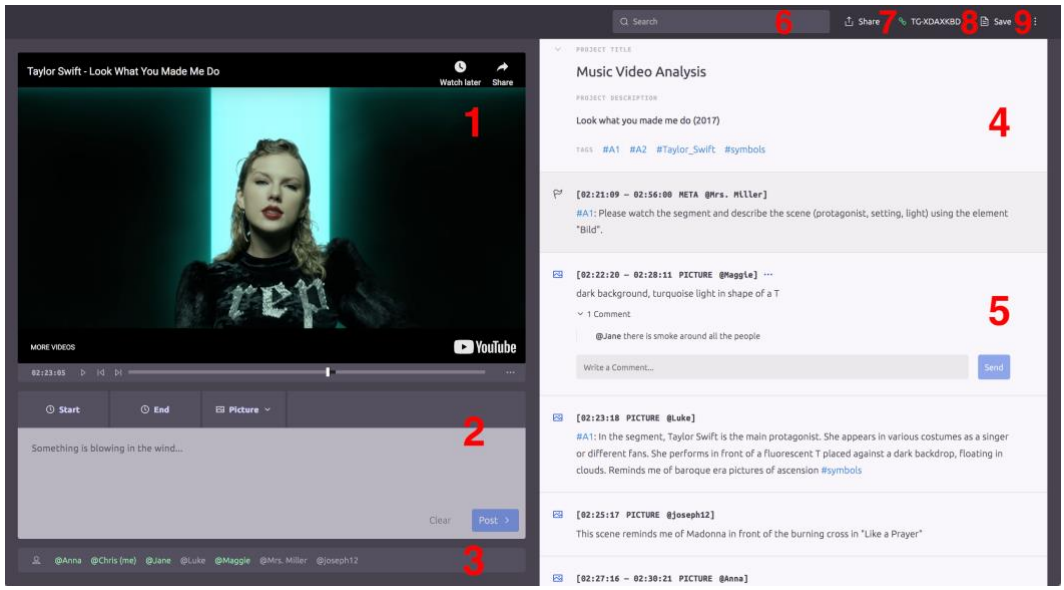


Figure 5. The TRAVIS GO work space provides a clear design and reduced features to match didactic requirements for teaching video annotation. The main work space areas are the video player (1), the text editor (2), and the annotation feed (4, 5).

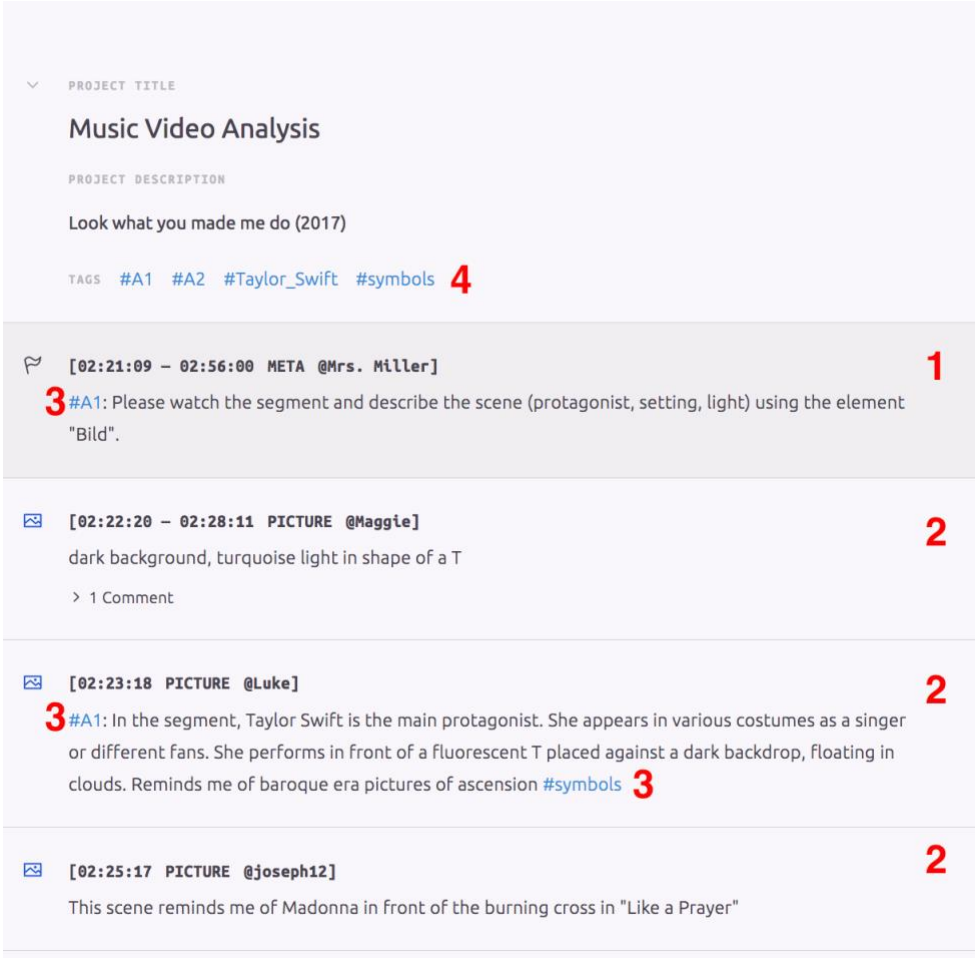


Figure 6. This detailed view of the annotation feed in TRAVIS GO shows the assignment posted by the teacher (1) and posts by students (2), the hashtags students to structure their answers (3) and the list of hashtags (4).

5.3.1 Tagging, filtering, editing, and commenting on posts

Figure 6 shows a more detailed view of the annotation feed in TRAVIS GO. When writing a post in the text editor, users can use the #-symbol to tag words (e.g. #house). Tags appear as list in the project description and as auto-complete suggestions in the text editor. Timestamp, category, username, and hashtags can be clicked to filter posts, multiple filters can be applied. Each post can be edited by the post author. Each user can comment on any post. Tags enable associative exploration and subsequent detailed analysis of audiovisual media artifacts. Users can use predefined tags and add new tags to structure their results. Tags (#) and user handles (@) also allow for references in annotating and provide teachers with a simple and efficient overview on students work. Filtering in the annotation feed allows to display posts by results or to review results of single students since multiple students are able to collaborate on one project together (see 5.4). Filters

help to evaluate progress and performance of individual students and learning results, for example, did students find and tag all scenes with a predefined tag. The basic interactive features in TRAVIS GO cater to curricula recommendations of how to teach about audiovisual media artifacts. They furthermore meet methodologic approaches teachers are already familiar with and improve the ways to work with audiovisual media artifacts.

5.3.2 Creating assignments and reviewing annotation results

To create assignments, teachers can use the meta category in the text editor to post one or more assignments in a project. They can define tags to indicate key words or analytical dimensions as part of assignments, for example, *"Describe the #camera_movement and #lighting in the opening scene"* or as presets for students to match with video sequences, for example, *"#closeup, #medium_shot, #long_shot"*. Filtering students' work can help teachers to consider results for grading classes. Teachers can assign students to comment on each other's posts to encourage peer-to-peer feedback, discussions or collaborative work which is a demand among teachers: *"Formal criteria are quite easy to teach, I guess, but how do I instruct students to reflect on it?"* (English teacher, Basel-City). Comments also enable teachers to give feedback with regard to students' individual contributions in group assignments.

5.4 Collaboration Mode

The analysis of teachers' educational needs and demands showed a strong demand for digital tools that allow better forms of collaborative work in peer-to-peer and student-to-teacher interaction. Teachers said, for example: *"We generally have difficulties in reaching the level of cooperative learning, going beyond working in groups. What I mean is collaborative knowledge development"* (Language teacher, Basel-City). TRAVIS GO supports collaborative work and exchange between students and teachers and students in a straightforward way (Klug, Schlote, and Eberhardt, 2017). The collaboration mode enables to initiate discussion in the web application and in face-to-face classroom interaction and allows teachers and students to give feedback on students' work. Multiple users can easily work simultaneously in the same project. Users generate a unique collaboration code (see Figure 7) that stays active as long as one user is active in the project. The collaboration code is displayed in the TRAVIS GO header (see Figure 8) and can be passed on by email, text etc., verbally, or by the teacher writing it on a board. Any person can join the project by entering the code into the app start page. This form of collaborating enhances existing teaching practices and adds didactic value to content-related depth, internal differentiation and cooperative learning. In this way, TRAVIS GO enables teacher-student inter-

action which is required and desired in curricula and by teachers to increase their general motivation to use a digital tool in teaching.

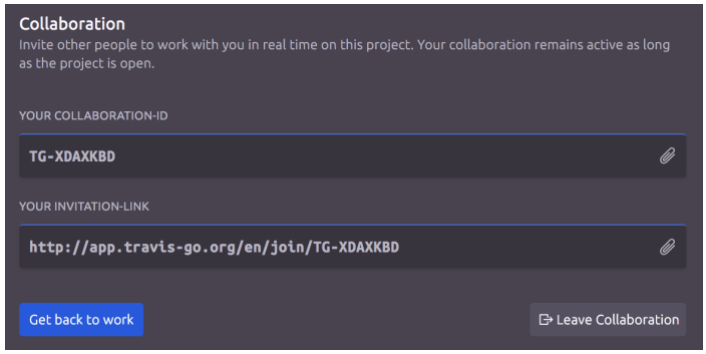


Figure 7. Any user can create a collaboration ID in a project at any time.

Figure 8. The collaboration ID can be shared via email, WhatsApp etc., verbally or by writing it on a board. The active collaboration ID is shown in the project menu bar.

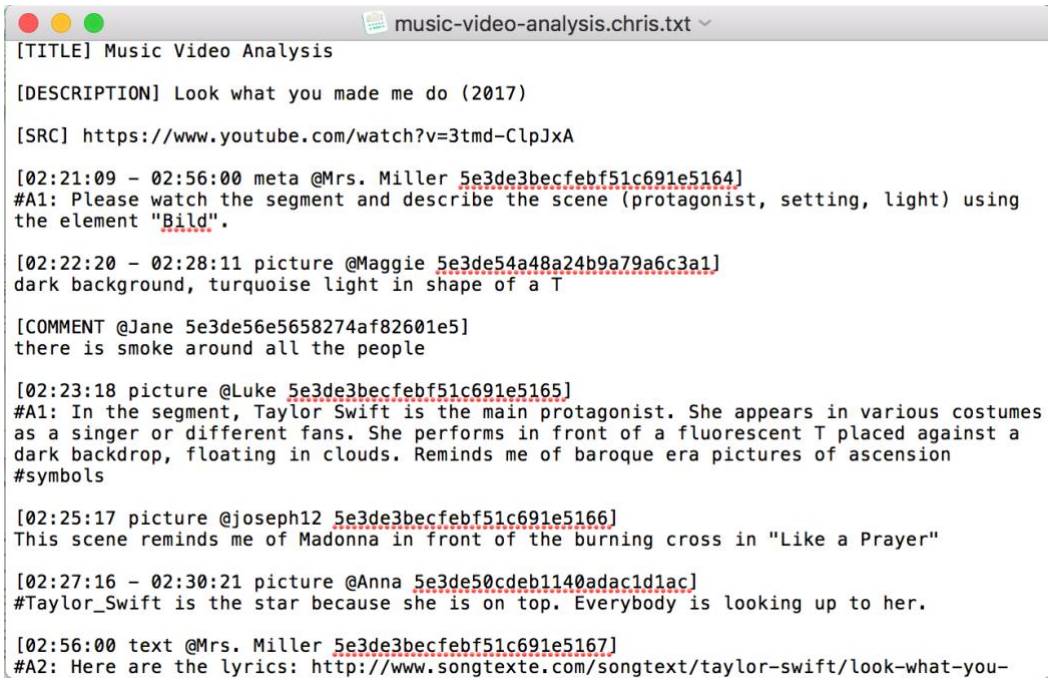


Figure 9. Any user in a project can save the project as text files.

5.5 Save and Share Projects

It is important, especially in educational contexts, that projects and analytical results can be easily exported from the web application into standard files. Because TRAVIS GO works without using a back-end data server, projects also need to be easily saved locally or in cloud services. In TRAVIS GO, users can save, load, and send projects as text files (see Figure 9). This meets schools' security standards using their own email and data servers. Users can also export projects as docx-files so students can create portfolios or hand them in as exams. Likewise, teachers can save projects as text files and share them with students or use them to start a collaboration for students to join (see 5.4).

6 Conclusion

We analyzed curricula for the school subjects *Languages*, *Music*, *Arts*, and *History*, and conducted 31 interviews with teachers in Switzerland and Germany. Curricula formulate the demand to teach analysis and interpretation of audiovisual media artifacts but do not give recommendations for adequate digital tools and provide only vague guidelines of how to integrate digital tools into existing teaching practices. From our interview study, we found the main problem to address are organizational limitations in teaching rather than the need for improved or more subject-specific analytical features or didactic environments. We found that an educational tool is much more beneficial for teachers when it is easy-to-use and fits into established teaching routines and that only basic analytical features are needed. The development of TRAVIS GO focused on designing a freely-available educational web application for simple and collaborate annotation of video and audio material that realizes these needs of teachers.

As a result, TRAVIS GO adds less value on methodological levels of educational video annotation tools in favor of openness and individual adaptability in subject-specific teaching contexts and by providing universal features for analyzing audiovisual media. The problem is that teachers are reluctant towards new digital tools if they are not comfortable using them, if they cannot adapt a tool into their teaching routines, and if a tool is too complex to easily use in little teaching time. TRAVIS GO solves these problems and adds greater value and innovation on organizational levels of integrating educational tools into established teaching routines. TRAVIS GO caters to teachers needs and concerns by reducing onboarding time, tool management, and time needed to familiarize oneself with its features and functionality.

- *TRAVIS GO provides an adequate analytical interface for educational use:*

The analysis of educational recommendations and educational needs and demands for video annotation tools revealed that teachers mainly use audiovisual media artifacts to illustrate facts and contexts (see 4.3) though they should teach analysis and interpretation (see 4.1). This is because of a lack of adequate tools and didactic approaches to perform more in-depth analytical discussions (see 4.5). TRAVIS GO supports this by providing an analytical interface that is action-oriented and allows exchange with others and to examine audiovisual media artifacts directly related to the material.

- *TRAVIS GO provides greater didactic freedom through reduced analytical structure:*

Teachers need an educational tool that provides subject-specific openness in creating assignments (see 4.3) and that can be integrated into established didactic settings without requiring additional effort from teachers (see 4.4). In TRAVIS GO, projects and assignments are not defined by the web application but can be designed according to subjects and learning goals to support a variety of didactic approaches. In order to offer added value beyond analog teaching-learning tools, digital learning tools should be customizable, interactive and adaptive. For TRAVIS GO, this has been realized by developing a reduced and open analytical structure that helps teachers to easily design learning activities around established didactic routines. TRAVIS GO is furthermore designed as a logical and consistent technical solution for didactic integration into existing teaching environments. This means, for example, it is up to teachers to choose audiovisual material, to assign tasks to groups or individuals or to decide to include students' contributions in TRAVIS GO into grading. Teachers as well need to supervise projects in TRAVIS GO and interact with students to coach them with their work.

TRAVIS GO was developed within educational context but is not limited to this purpose. We pointed out that our focus is to provide an easy-to-use web app that allows for a quick and straightforward integration into inevitable limitations of existing educational situations and not on designing new subject-specific analytical features or environments. Accessibility, usability, and adaptability to existing practices are more crucial keys for deciding to use a digital tool. Analytical dimensions and features are universal for analyzing and annotating any audiovisual format in any context and differ in their interpretative perspectives on each subject and context. TRAVIS GO provides a free workspace for individual specification based on universal analytical dimensions. This allows to use TRAVIS GO in any other context that involves forms of collaboratively reviewing, discussing, or annotating audiovisual media artifacts, such as video production, video evaluation, or higher education.

Overall, our results contribute further insights for the CSCW and the CSCL community about the need to include user demands into developing educational tools and how to implement these in the development and design process of a web application for collaborative video annotation. Schools and teachers are main agents in supporting students' media literacy. Teenagers frequently consume and interact with audiovisual media artifacts. But teachers lack adequate didactic tools and methodological skills to analyze audiovisual media artifacts according to curricula standards. TRAVIS GO demonstrates a successful way how to solve these issues and challenges in the design and development of an open educational resource (OER).

7 Limitations and Future Work

Our study is limited to that we did not evaluate students' needs and demands as part of designing TRAVIS GO. This is because we identified teachers as more crucial regarding openness, prejudices, and knowledgeability towards introducing and using new tools into existing teaching routines, therefore is it important to primarily evaluate teachers' attitudes, mindsets, and ideologies when designing educational tools. Our results are furthermore limited because we are not including data about the validation of TRAVIS GO in realistic educational environments. Although we tested and evaluated TRAVIS GO through video observations of students and teachers working with the app, data collection from using TRAVIS GO in various school subjects and didactic settings is still ongoing and part of future work.

References

Ahuja, R., D. Khan, D. Symonette, M. des Jardins, S. Stacey, and D. Engel (2019): 'A Digital Dashboard for Supporting Online Student Teamwork'. In: *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. pp. 132–136.

Al-Zaidiyeen, N. J., L. L. Mei, and F. S. Fook (2010): 'Teachers' Attitudes and Levels of Technology Use in Classrooms: The Case of Jordan Schools'. *International education studies*, vol. 3, no. 2, pp. 211–218.

Alper, B., N. H. Riche, F. Chevalier, J. Boy, and M. Sezgin (2017): 'Visualization Literacy at Elementary School'. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA, pp. 5485–5497.

Anderson, M. and J. Jiang (2018): 'Teens, social media & technology 2018'. *Pew Research Center*, vol. 31, pp. 2018.

Arpaci, I., K. Kilicer, and S. Bardakci (2015): 'Effects of security and privacy concerns on educational use of cloud services'. *Comput. Human Behav.*, vol. 45, pp. 93–98.

Baecker, R. M., D. Fono, and P. Wolf (2007): 'Towards a video collaboratory'. *Video research in the learning sciences*, pp. 461–478.

Barab, S. and K. Squire (2004): 'Design-Based Research: Putting a Stake in the Ground'. *Journal of the Learning Sciences*, vol. 13, no. 1, pp. 1–14.

Bargeron, D., J. Grudin, A. Gupta, E. Sanocki, F. Li, and S. Leetiernan (2002): 'Asynchronous collaboration around multimedia applied to on-demand education'. *Journal of Management Information Systems*, vol. 18, no. 4, pp. 117–145.

Erziehungsdepartement Basel-Stadt (2013): 'LEHRPLAN GYMNASIUM Übergangszeit 2014 – 2021'. URL: https://www.edubs.ch/unterricht/lehrplan/mittelschulen/dokumente/downloads/lehrplan-gymnasien-uebergangszeit-2014-2021 [18.02.2021]

BMBWF (2019): 'Bildnerische Erziehung'. URL: https://bildung.bmbwf.gv.at/schulen/unterricht/lp/ahs4_778.pdf?61ebzm. [18.02.2021]

Bogner, A., B. Littig, and W. Menz (2009): 'Introduction: Expert interviews—An introduction to a new methodological debate'. In: *Interviewing experts*. Springer, pp. 1–13.

Bordwell, D., Thompson, K., & Smith, J. (2017). *Film art: An introduction*. New York: McGraw-Hill.

Bundes Jugend Vertretung (2017): '#MeinNetz - Internetnutzung & Medienkompetenz junger Menschen in Österreich'. Technical report, Wien, Austria. URL: http://mein-netz.at/studie-meinnetz-internetnutzung-medienkompetenz-junger-menschen-in-oesterreich/ [18.02.2021].

Bundesamt für Sozialversicherung (2019): 'Medienbildung – eine Herausforderung für die Schule'. URL: https://www.jugendundmedien.ch/medienkompetenz-foerdern/lehrpersonen-schule.html [18.02.2021].

Burr, B. (2006): 'VACA: A Tool for Qualitative Video Analysis'. In: *CHI '06 Extended Abstracts on Human Factors in Computing Systems*. New York, NY, USA, pp. 622–627.

Cadiz, J. J., A. Gupta, and J. Grudin (2000): 'UsingWeb annotations for asynchronous collaboration around documents'. In: *Proceedings of the 2000 ACM conference on Computer supported cooperative work*. pp. 309–318.

Carneiro, G., M. Nacenta, A. Toniolo, G. Mendez, and A. J. Quigley (2019): 'Deb8: A Tool for Collaborative Analysis of Video'. In: *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*, pp. 47–58.

Chen, D. l., D. Freeman, and R. Balakrishnan (2019): 'Integrating Multimedia Tools to Enrich Interactions in Live Streaming for Language Learning'. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 438:1–438:14.

Chien, S.-P., H.-K. Wu, and Y.-S. Hsu (2014): 'An investigation of teachers' beliefs and their use of technology-based assessments'. *Comput. Human Behav.*, vol. 31, pp. 198–210.

Cobb, P., J. Confrey, A. diSessa, R. Lehrer, and L. Schauble (2003): 'Design Experiments in Educational Research'. *Educ. Res.*, vol. 32, no. 1, pp. 9–13.

Crabtree, A., T. Rodden, and J. Mariani (2004): 'Collaborating around collections: informing the continued development of photoware'. In: *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. pp. 396–405.

Cunha, B. C. R., O. J. Machado Neto, and M. d. G. Pimentel (2013): 'MoViA: A Mobile Video Annotation Tool'. In: *Proceedings of the 2013 ACM Symposium on Document Engineering*, pp. 219–222.

Design-Based Research Collective (2003): 'Design-Based Research: An Emerging Paradigm for Educational Inquiry'. *Educ. Res.*, vol. 32, no. 1, pp. 5–8.

Erziehungsdirektion des Kantons Bern (2017): 'Lehrplan 17 für den gymnasialen Bildungsgang'. URL: https://www.erz.be.ch/erz/de/index/mittelschule/mittelschule/gymnasium/lehrplan_maturitaetsausbildung.assetref/dam/documents/ERZ/MBA/de/AMS/GYM20LP2017/ams_gym_lehrplan_2017_gesamtdokument.pdf. [18.02.2021].

Diakopoulos, N., S. Goldenberg, and I. Essa (2009): ‘Videolyzer: Quality Analysis of Online Informational Video for Bloggers and Journalists’. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 799–808.

Diamant, E. I., S. R. Fussell, and F.-l. Lo (2008): ‘Where did we turn wrong? Unpacking the effect of culture and technology on attributions of team performance’. In: *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. pp. 383–392.

Dodson, S., I. Roll, M. Fong, D. Yoon, N. M. Harandi, and S. Fels (2018): ‘An Active Viewing Framework for Video-based Learning’. In: *Proceedings of the Fifth Annual ACM Conference on Learning at Scale*, pp. 24:1–24:4.

Dorn, B., L. B. Schroeder, and A. Stankiewicz (2015): ‘Piloting TrACE: Exploring spatiotemporal anchored collaboration in asynchronous learning’. In: *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pp. 393–403.

El Raheb, K., A. Kasomoulis, A. Katifori, M. Rezkalla, and Y. Ioannidis (2018): ‘A Web-based System for Annotation of Dance Multimodal Recordings by Dance Practitioners and Experts’. In: *Proceedings of the 5th International Conference on Movement and Computing*, pp. 8:1–8:8.

Flick, U. (2018): *An introduction to qualitative research*. New York, Sage.

Frees, B., T. Kupferschmitt, and T. Müller (2019): ‘ARD/ZDF-Massenkommunikation Trends 2019: Non-lineare Mediennutzung nimmt zu’. *Media Perspektiven*, vol. 2019, no. 7–8, pp. 314–333.

Hamilton, W. A., N. Lupfer, N. Botello, T. Tesch, A. Stacy, J. Merrill, B. Williford, F. R. Bentley, and A. Kerne (2018): ‘Collaborative Live Media Curation: Shared Context for Participation in Online Learning’. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 555:1–555:14.

Hartmann, B., M. R. Morris, H. Benko, and A. D. Wilson (2010): ‘Pictionaire: supporting collaborative design work by integrating physical and digital artifacts’. In: *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. pp. 421–424.

Heimonen, T., O. Koskinen, J. Okkonen, R. Raisamo, M. Turunen, S. Kangas, T. Pallos, P. Pekkala, S. Saarinen, K. Tiitinen, T. Keskinen, and M. Luhtala (2013): ‘Seek’N’Share’ A platform for location-based collaborative mobile learning. In *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*, pp. 1-4.

Hickey, S. (2019): ‘Bricoleur: A Tool for Tinkering with Programmable Video and Audio’. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. LBW0118:1–LBW0118:6.

Holstein, K., B. M. McLaren, and V. Aleven (2017a): ‘Intelligent Tutors As Teachers’ Aides: Exploring Teacher Needs for Real-time Analytics in Blended Classrooms’. In: *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pp. 257–266.

Hong, L. and E. H. Chi (2009): ‘Annotate Once, Appear Anywhere: Collective Foraging for Snippets of Interest Using Paragraph Fingerprinting’. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1791–1794.

Hosack, B. (2010): ‘VideoANT: Extending online video annotation beyond content delivery’. *TechTrends*, vol. 54, no. 3, pp. 45–49.

International Society for Technology in Education (ISTE) (2017): ‘Tools to Support Digital Learning’. URL: https://www.iste.org/explore/In-the-classroom/5-open-educational-resources-to-try [18.02.2021]

Jiang, H., A. Viel, M. Bajaj, R. A. Lue, and C. Shen (2009): ‘CThru: Exploration in a Video-centered Information Space for Educational Purposes’. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1247–1250.

Kim, J., S.-W. Li, C. J. Cai, K. Z. Gajos, and R. C. Miller (2014): 'Leveraging video interaction data and content analysis to improve video learning'. In: *Proceedings of the CHI 2014 Learning Innovation at Scale workshop*.

Klug, D., and E. Schlote (2018): 'Ästhetische Bildung mit audiovisuellen Medien digital unterstützen - schulischer Praxisbedarf und Konzepte der Filmbildung'. In: U. Autenrieth, D. Klug, A. Schmidt, and A. Deppermann (eds.): *Medien als Alltag*. Köln: Herbert von Halem, pp. 68–98.

Klug, D., Schlote, E., & Eberhardt, J. O. (2017). 'Musikvideos im Fremdsprachenunterricht – Wie der Einsatz einer Web-Applikation Binnendifferenzierung und kooperatives lernen ermöglicht'. In: *Babylonia, 26*(3), pp. 34–37.

Kumar, P. C., Chetty, M., Clegg, T. L., & Vitak, J. (2019). Privacy and security considerations for digital technology use in elementary schools. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1-13.

Lange, P. G. (2016): *Kids on YouTube: Technical identities and digital literacies*. New York, Routledge.

Liebau, E., L. Klepacki, and J. Zirfas (2009): '*Theatrale Bildung'. Theaterpädagogische Grundlagen und kulturpädagogische Perspektiven für die Schule*. Weinheim/München, Juventa.

Lim, S. and P. Chiu (2015): 'Collaboration Map: Visualizing temporal dynamics of small group collaboration'. In: *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing*, pp. 41–44.

Liu, C., J. Kim, and H.-C. Wang (2018): 'ConceptScape: Collaborative Concept Mapping for Video Learning'. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 387:1–387:12.

Liu, C. j., C.-L. Yang, J. J. Williams, and H.-C. Wang (2019): 'NoteStruct: Scaffolding Note-taking While Learning from Online Videos'. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. LBW1116:1–LBW1116:6.

Marlow, J., S. Carter, N. Good, and J.-W. Chen (2016): 'Beyond talking heads: multimedia artifact creation, use, and sharing in distributed meetings'. In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pp. 1703–1715.

McDonald, N., S. Schoenebeck, and A. Forte (2019): 'Reliability and inter-rater reliability in qualitative research: Norms and guidelines for CSCW and HCI practice'. *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–23.

McNally, J. and B. Harrington (2017): 'How Millennials and Teens Consume Mobile Video'. In: *Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*, pp. 31–39.

Medienpädagogischer Forschungsverbund Südwest (2018): 'JIM-Studie. Basisuntersuchung zum Medienumgang 12-19-Jähriger'. URL: https://www.mpfs.de/studien/jim-studie/2018/ [18.02.2021].

Merz, A., A. Hu, and T. Lin (2018): 'ClipWorks: A Tangible Interface for Collaborative Video Editing'. In: *Proceedings of the 17th ACM Conference on Interaction Design and Children*, pp. 497–500.

Meuser, M. and U. Nagel (2009): 'The Expert Interview and Changes in Knowledge Production'. In: A. Bogner, B. Littig, and W. Menz (eds.): *Interviewing Experts*. London: Palgrave Macmillan, pp. 17–42.

Ministrium für Kultus Jugend und Sport Baden-Württemberg, M. (2016): ‘Bildungsplan 2016 – Bildungsplan der Oberstufe an Gemeinschaftsschulen: Bildende Kunst’. URL: www.bildungsplaene-bw.de/site/bildungsplan/get/documents/lsbw/export-pdf/depot-pdf/ALLG/BP2016BW_ALLG_GMSO_BK.pdf [18.02.2021].

Nguyen, P., J. Kim, and R. C. Miller (2013): ‘Generating Annotations for How-to Videos Using Crowdsourcing’. In: *CHI ’13 Extended Abstracts on Human Factors in Computing Systems*, pp. 835–840.

Pollalis, C., C. Grevet, L. Westendorf, S. Finn, O. Shaer, and P. Metaxas (2018): ‘Classroom Activity for Critical Analysis of News Propagation Online’. In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. CS05:1–CS05:10.

Rideout, V. J. and M. B. Robb (2019): *The common sense census: Media use by tweens and teens*. Common Sense Media.

Rohleder, D. B. (2019): ‘Smart School –Auf dem Weg zur digitalen Schule’. URL: https://www.bitkom.org/sites/default/files/2019-03/Pr{\"a}4sentationBitkom-PKBildungskonferenz 2012.03.2019_final.pdf, month =mar [20.05.2020].

Sang, G., M. Valcke, J. van Braak, J. Tondeur, and C. Zhu (2011): ‘Predicting ICT integration into classroom teaching in Chinese primary schools: exploring the complex interplay of teacher related variables’. *Journal of Computer Assisted Learning*, vol. 27, no. 2, pp. 160–172.

Schlote, E., and D. Klug (2020): ‚Ein digitales Lernwerkzeug realisieren – der Entwicklungsprozess der Web-Applikation TRAVIS GO an der Schnittstelle von Medienwissenschaft, Informatik und Schulpädagogik’. In: *Schnittstellen und Interfaces – Digitaler Wandel in Bildungseinrichtungen, 7*, pp. 169–185.

Schlote, E., D. Klug, and K. Neumann-Braun (2020): ‚Mittendrin statt nur dabei: Partizipation im schulischen Unterricht mit der Web-App TRAVIS GO digital unterstützen’. In: *MedienPädagogik: Zeitschrift für Theorie und Praxis der Medienbildung*, pp. 507–529.

Serrano-Cámara, L. M., M. Paredes-Velasco, C. M. Alcover, and J. Á. Velazquez-Iturbide (2014): ‘An evaluation of students’ motivation in computer-supported collaborative learning of programming concepts’. In: *Computers in human behavior, 31*, pp. 499–508.

Shi, Y., Y. Wang, Y. Qi, J. Chen, X. Xu, and K.-L. Ma (2017): ‘IdeaWall: Improving creative collaboration through combinatorial visual stimuli’. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. pp. 594–603.

Shin, H., E.-Y. Ko, J. J. Williams, and J. Kim (2018): ‘Understanding the Effect of In-Video Prompting on Learners and Instructors’. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, p. 1-12.

Singh, V., S. Abdellahi, M. L. Maher, and C. Latulipe (2016): ‘The Video Collaboratory As a Learning Environment’. In: *Proceedings of the 47th ACM Technical Symposium on Computing Science Education*, pp. 352–357.

Singh, V., Latulipe, C., Carroll, E., & Lottridge, D. (2011, November). The choreographer's notebook: a video annotation system for dancers and choreographers. In *Proceedings of the 8th ACM Conference on Creativity and Cognition*, pp. 197-206.

Suter, L., G. Waller, J. Bernath, C. Külling, I. Willemse, and others (2018): ‘JAMES: Jugend, Aktivitäten, Medien-Erhebung Schweiz’. URL: https://www.zhaw.ch/de/psychologie/forschung/medienpsychologie/mediennutzung/james/#c157373 [18.02.2021].

Vega, V. and M. Robb (2019): *The Common Sense census: Inside the 21st-century classroom*. San Francisco, CA: Common Sense Media.

Wang, F. and M. J. Hannafin (2005): ‘Design-based research and technology-enhanced learning environments’. *Educ. Technol. Res. Dev*., vol. 53, no. 4, pp. 5–23.

Wang, I., P. Narayana, J. Smith, B. Draper, R. Beveridge, and J. Ruiz (2018): 'EASEL: Easy Automatic Segmentation Event Labeler'. In: *23rd International Conference on Intelligent User Interfaces*, pp. 595–599.

Weng, C. and J. H. Gennari (2004): 'Asynchronous collaborative writing through annotations'. In: *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pp. 578–581.

Yim, S., D. Wang, J. Olson, V. Vu, and M. Warschauer (2017): 'Synchronous Collaborative Writing in the Classroom: Undergraduates' Collaboration Practices and their Impact on Writing Style, Quality, and Quantity'. In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 468–479.

Yoon, D., N. Chen, B. Randles, A. Cheatle, C. E. Löckenhoff, S. J. Jackson, A. Sellen, and F. Guimbretière (2016): 'RichReview++ Deployment of a Collaborative Multi-modal Annotation System for Instructor Feedback and Peer Discussion'. In: *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pp. 195–205.

Qingxiaoyang Zhu, Hao-Chuan Wang (2021): Is a GIF Worth a Thousand Words? Understanding the Use of Dynamic Graphical Illustrations for Procedural Knowledge Sharing on wikiHow. In: Proceedings of the 19th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies - Exploratory Papers, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591), DOI: 10.18420/ecscw2021_ep16

Is a GIF Worth a Thousand Words? Understanding the Use of Dynamic Graphical Illustrations for Procedural Knowledge Sharing on wikiHow

Qingxiaoyang Zhu, Hao-Chuan Wang
University of California, Davis
qinzhu@ucdavis.edu, hciwang@ucdavis.edu

Abstract. Informational presentations in Graphics Interchange Format (GIF), have been commonly used to convey emotional, cultural, and non-verbal cues in computer-mediated communication. However, the usage and impact of animated GIFs in the sharing and consumption of procedural knowledge, such as how-to instructions, remains unclear. In this paper, we leverage an online collaborative procedural knowledge-sharing platform – wikiHow to investigate the roles of GIFs in the construction and editing of How-To tutorials and how multimodal tutorials impact learners' perception and learning from the shared expositions. Through data analytics of archived editing histories, article content and user ratings of wikiHow pages, we found that tutorials consisting of multiple modalities, including animated GIFs and images in addition to text, in general introduced more edits and more textual content compared with text-only articles. When learners learned from these wikiHow tutorials, the tutorials with rich modalities also received higher usefulness evaluation from learners, and accumulated more success stories from the learners in following the tutorials to perform procedural tasks. We discuss the implications for future research on multimodal factors for collaborative procedural knowledge sharing.

Introduction

With the advent of the Internet and computer-mediated communication (CMC) technologies, collaborative knowledge sharing is rising on digital platforms such as Wikipedia, where people distributed around the world may collaborate to make contributions to encyclopedic repositories to provide access to diverse topical knowledge and concepts (Ackerman et al. (2013)). One emerging trend in online knowledge sharing targets at leveraging multimodal information, such as images and videos along with text, to support readers' understanding of the topic. Numerous research work has investigates the integration of text and images in documents with the goal to relieve the cognitive effort for readers to comprehend the content (Viegas (2007); Navarrete and Villaespesa (2020)). However, as knowledge sharing requires authors to actively create expositions to share what they know, how the inclusion of multimodal information affects the construction process of knowledge sharing, such as editing and peer collaboration for producing an exposition useful to learners, is important yet not entirely clear.

Apart from vast knowledge on diverse topics shared by people online today, instructions of procedural skills that teach learners how to accomplish specific tasks are highly practical and frequently needed in daily life. Such procedural knowledge gains increasing popularity in the landscape of knowledge management and emerges as a new genre of online content of increasing demand by online communities to share and to consume. Specifically, How-To instructions refer to procedural knowledge externalized and authored by people that explains how to achieve a desired goal through a series of operational steps (Torrey et al. (2007); Yang and Wang (2019)). For instance, how to fix a disposal, how to bake a Chiffon cake, etc.. Similar to Wikipedia, How-Tos are written and edited by volunteers (wikiHow is one popular platform for authors to share How-Tos; see Figure 1). In comparison with traditional face-to-face knowledge transfer that tends to have limited scalability, volunteering-based online knowledge sharing transcends

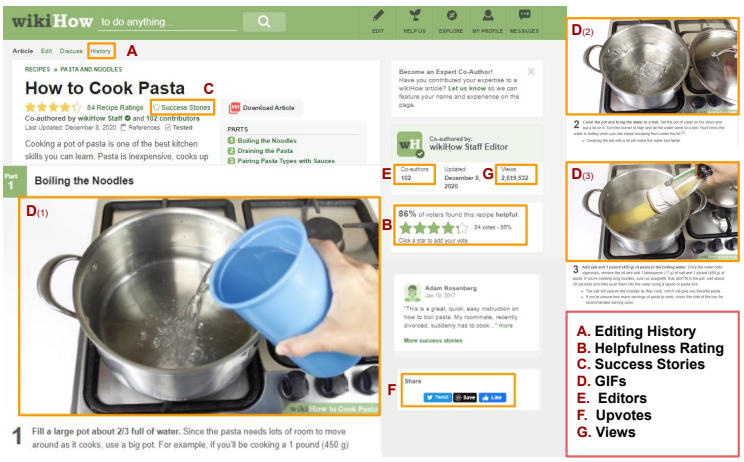


Figure 1: wikiHow as a collaborative procedural knowledge sharing platform.

geographic and social boundaries, enabling the curation of procedural contents from volunteers at a large scale. However, as volunteers don't necessarily possess domain knowledge to the same level, the knowledge gap between editors can potentially cause difficulties in coordination and collaboration during the editing process of How-To instructions, as shown in previous research of Wikipedia by Shaw and Hargittai (2018). Although a lot of work has been done on supporting collaborative editing among volunteers on Wikipedia (Liu and Ram (2011); Brandes et al. (2009)), few work have been done on online procedural knowledge sharing.

Different from knowledge sharing on Wikipedia that tends to focus on the sharing of declarative knowledge, the sharing of procedural knowledge requires editors to externalize the procedural skills they possess, which is in theory more difficult to verbalize and write about than editing regular encyclopedic entries (Yang (2007)). Also, experts may not be able to fully articulate their how-to procedures in a form that novices can understand (Hinds et al. (2001)), as they may unintentionally describe the procedures with abstract forms, personal metaphors or terminologies. Also, what's articulated by one editor may also be hard to be further edited by other editors with inter-editor discrepancies on skills and expertise, which may all hinder procedural knowledge sharing.

On the other hand, sharing procedural knowledge versus sharing conceptual knowledge, as in traditional Wikipedia, may differ in what's the feasible modality of informational presentation to use for exposition and conveyance. Different from conceptual knowledge, procedural knowledge involves also psychomotor skills and operating behaviors (Salaberry (2018)), where images and GIFs are potentially useful for explicating and communicating these skills. The visual representations provide the affordances essential to represent procedural behaviors in instructional forms for learners. In terms of inter-editor collaboration, previous work (Gergle et al. (2013)) has suggested visual information benefits collaborative work by grounding inter-worker communication. Similar visualization effects may also benefit collaborative editing in procedural knowledge sharing given the non-verbal nature of the procedural content to co-author and to share. Based on these observations, we're interested in investigating how multimodal representations (animated GIFs, static images and texts) are used to facilitate how-to instruction authoring among editors and procedural knowledge sharing from experts to novices in the field, such as the online how-to knowledge sharing platform, wikiHow as shown in Fig. 1.

Beyond static images, dynamic graphical representations, such as video and GIFs, have the affordances to illustrate time-dependent processes, such as physical actions and object states visually and can potentially complement other modalities (like text and images) to better convey procedural knowledge. From the learning perspective, inferring the underline meaning of messages conveyed by visual channels may require learners to jointly interpret information coming both from text and graphics, leading to synergistic effects of multimodal learning (Marsh and White (2003)). What remains underexplored is the effects of dynamic graphics (e.g., GIFs) and mixed pictorial representations (e.g., a combination of static images

and GIFs) in procedural knowledge sharing, where text-only expositions can be insufficient for both knowledge sharers and learners given the non-verbal and abstract nature of procedural knowledge (Volker et al. (2003); Hinds et al. (2001)).

In order to gain deeper understanding on how multimodal procedural knowledge sharing occurs by combining text, static images and dynamic graphical representations, and how multimodal instructions created impact novice learners' evaluations of how-to contents, we leverage the online platform wikiHow[1], an community that shares and manages how-to knowledge, to understand how readers' evaluations of tutorials and learning outcomes vary across the modalities adopted to expose procedural knowledge at a large scale.

Background

Graphics Interchange Format (GIF) is a graphic file format increasingly popular in communication and social networking applications for encoding and exchanging animating graphical contents (Bakhshi et al. (2016)). Similar to pictorial emojis, emoticons and stickers, short videos encoded in GIFs are also commonly used as emotional displays, such as animating gestures and facial expressions. The polysemic and intertextual features of GIFs make such format feasible to convey rich personalized expressions, affects and cultural knowledge (Miltner and Highfield (2017); Tolins and Samermit (2016)). Previous works on GIFs mostly focus on investigating the affordances of GIFs in online communication, such as social interactions in social media posts and conversations in instant messages (Jiang et al. (2018)). The utilities of GIFs in the context of knowledge sharing and consumption remains unclear.

Procedural Knowledge Transfer is another line of study aligned with our research. In previous research, Hinds et al. (2001) identified the difficulty during knowledge transfer between experts and novices, where the abstraction level of expertise may block such knowledge transfer. Previous work by Huang and Chiou (2010) also investigates how different media included in the instructions influence the process and outcomes of completing procedural tasks. The result reveals that alternative visual types for instruction should be utilized to facilitate learning. On the other hand, Chirumalla et al. (2015) stated people using text-only instructions for knowledge sharing may take three times longer to accomplish the task than other instructions, while Palmiter et al. (1991) claimed that animated demonstrations performed worse in supporting transfer learning. Therefore, it is important to investigate how to combine different modalities to support effective and efficient knowledge sharing for procedural tasks.

[1] wikiHow Introduction: https://www.wikihow.com/wikiHow:About-wikiHow, Access Date: 10.18.2020

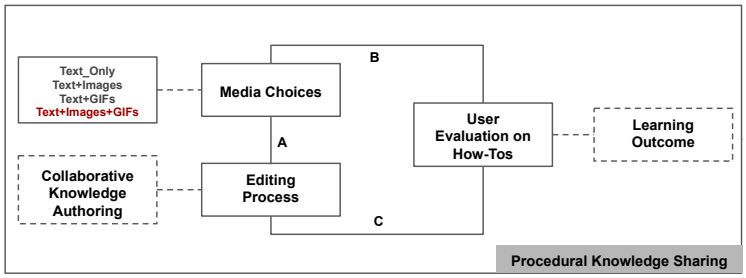


Figure 2: The analysis framework: the framework for analyzing how media modalities affect collaborative editing by editors and learners' ratings of the contents created. The articles are labeled into four categories: text-only, text+images, text+GIFs, text+images+GIFs, according to the media used in the editing history. The correlation between media choices and the collaborative editing process (linkA) is examined. Besides, the connection between media choices, the editing process and the learners' evaluation is investigated (link B & C).

Hypotheses and Analysis Framework

We focus on analyzing all English tutorials posted and shared on wikiHow. We collect two sets of data, including (1) the editing history of tutorials reflecting the behaviors of collaborative authoring by editors and (2) the metadata revealing viewers' interaction behavior when they acquire procedural skills on the platform, for instance, viewers' helpfulness rating and upvotes on articles, success stories and etc..

To examine the roles of multimodal representations in collaborative knowledge sharing for how-to knowledge, we pose the following two hypotheses:

H1: Compared with non-GIF incorporated instructions of procedural knowledge, GIF incorporated instructions trigger more editing behaviors and elicit more textual contents.

H2: Compared with non-GIF incorporated representations of procedural knowledge, GIF incorporated instructions better support regular learners' learning.

To examine aforementioned hypotheses H1 and H2, we follow the analysis framework illustrated by Fig. 2. The framework models the main components that happen during collaborative procedural knowledge sharing, namely collaborative knowledge authoring and novice learners' knowledge consumption, where our focus lies in the middle on how media choices correspond to collaborative knowledge sharing. In the figure, *Link A* represents the potential connection between media choices and collaborative editing activities, while *Link B* and *Link C* indicates the potential connection between media types included in the instructions, edits from editors and the evaluation from novice learners on the instructions. We follow *Link A* to test H1, and follow *Link B, C* to check H2.

Method and Data

To investigate what the editing behaviors and editing outcomes were by volunteer editors when dynamic graphics in GIFs are included in how-to tutorials for knowledge conveyance, as well as viewers' perception and evaluation on how-to instructions using different modalities, we extracted online How-Tos instructions covering various topics and their metadata from wikiHow to conduct statistical analyses of archived data. The metadata included the editing history of each instruction and the viewers' interaction data, such as viewers' helpfulness rating, upvotes and shared successful stories. Each How-To instruction was initialized by an expert editor who wants to share the skill to accomplish a task, then the instruction was edited by multiple volunteers. The viewers were allowed to evaluate how helpful a tutorial was and share their successful experience (i.e., success stories following the tutorial) after consuming the instructions.

With the large-scale data collected, we were able to explore the behaviors of collaborative editing and knowledge sharing in the field. In order to model the influence of graphical representation, four media types were used to encode each tutorial instruction according to the media modalities involved in its editing history: *text only, text+images, text+GIFs text+images+GIFs*. When encoding media types used in tutorials, we carefully cleaned the data by removing noisy edits of multimedia contents, including the insertion of apparently irrelevant images that survived only for one edit (reverted by the following edit) in the editing history. Further, the *number of edits* along history and the *number of words* in the last-version of each instruction are calculated. The purpose is to quantify the effort paid by editors to author the procedural instructions and the externalized information conveyed by the accompanying text respectively. To deal with the confounding effect of the tutorial length and the tutorial age (e.g., an old tutorial usually tends to have more edits and a long tutorial usually contains more textual contents), we carefully normalize the word counts according to the number of steps included by the tutorial. For the edits number, the similar long-tail phenomenon of edits is observed inside time slices as previous work demonstrates (Wilkinson and Huberman (2007)). Therefore, we adopt the similar method to split instructions into four equal-sized time slices according to article ages, the number of articles is equal for each time slice. Then, we log-normalize and time-normalize edits inside the slice along time. Finally, we extract the features related to the viewers, including their *rating on helpfulness* and *the number of success stories* following a tutorial posted by viewers. The reason to extract these ratio numbers is that they are the lightweight signals reflecting novices' evaluation of instructions and learning outcome. In addition, they are measurable signals indirectly revealing viewers' interpretation results. Besides, the number of successful stories is normalized according to the tutorial age to eliminate the legacy effect that the number is accumulative along the time.

In total, the collected dataset covers the archive of 249,465 unique English How-To articles over the past 16 years on the wikiHow platform. These articles

are categorized into 19 main topics (Hobbies and Crafts, Personal Care and Style, Food and Entertaining etc.). 110,694 out of 249,465 (44.37%) pages are marked as stubs in overall due to short length and insufficiently conveying enough information. These stub pages and pages in the quality review process are not included, because low quality pages do not receive feedback from viewers and are not valid for studying collaborative editings. GIFs' usage is of our interest, we therefore count the number and proportion of the articles that include GIFs against the total number of articles of each main topic. GIFs are used in all topic categories, among which "Personal Care and Style" (2,165 out of 11,194) and "Food and Entertaining" (2,501 out of 12,580) topics contain GIFs with the highest proportion (more than 20%). As such, we mostly focus on these two topics for analysis, so that abundant data is available and generalizability exists. Because of an unbalanced distribution across the four media encodings, we apply stratified sampling to the data to reconstruct a representative subset consisting of 1168 samples for empirical data analysis. By doing these steps, we carefully handle the article quality and topic variation.

Preliminary Results

We report the intermediate analysis results and key findings of archived How-To tutorials in this section. The findings are from two perspectives: volunteers' editing activities of tutorial instructions, and viewers' perception and evaluation on these instructions.

Analysis of Knowledge Production

To address the questions of how combinations of different modalities used in tutorials affect the effectiveness of knowledge sharing, we first examined underlying patterns associated with media modalities during the time editors authoring the tutorials. The editing process was of our interest because the tutorial presented to the novices was the production of multi-party work. The editing behaviors directly determined the content of the tutorial, and further influenced the viewers' perception and evaluation. In our preliminary analysis, the normalized number of edits were used to measure editing activities, and normalized word counts were used to measure the editing outcomes.

Media Choices and Editing Behaviors

As How-To instructions on wikiHow evolve all the time, no final versions are available. Our analysis thus could simply focus on the last snapshot of each tutorial by controlling article age and topic statistically as shown in the previous section. The number of edits was first log normalized and then time-normalized, so that a fair comparison can be achieved. To test how editing activities differ between articles that involve different combinations of modalities, we conducted ANOVA and post-hoc Tukey HSD tests by using media modalities used in an article as the independent

variable and normalized number of edits as the dependent variable. The topic was set as a control variable. The result showed a main effect of media modalities used on normalized edit times, $F(3, 1164) = 22.4352, p < 0.0001, \eta = 0.05$. Significant differences were found between rich media combination (including both dynamical GIFs and static images to complement text) and every other media combinations, as well as the text-only baseline (all *p*s<.05), where more edits appeared when GIFs and images were both added to the tutorials at some points. It was worth noting that including only images or only GIFs did not correspond to increasing edits compared with text-only baseline. Besides, introducing GIFs alone was still associated with a significantly higher number of edits compared to introducing static images only. The results revealed that integrating dynamic and static visual information in tutorials is a significant predictor of more productive editing behaviors by the knowledge sharers. There could be either a potential image-elaboration effect where the inclusion of rich media content motivates authors to contribute more edits to elaborate the images/animations, thus potentially share also more and better procedural knowledge; or a potential text-illustration effect where edits to tutorials motivate the needs to include visual representations to clarify the procedures.

Media Choices and Editing Outcomes

In order to quantitatively measure the editing outcomes, we used word counts to approximate the quantity of externalized information conveyed by the accompanying text. Since stop words, such as the most common and short function words (the, a, is...), provided little information, they were filtered out. Procedural instructions to accomplish a task naturally consists of a series of steps. The methods to complete a task could also vary from person to person, thus How-To instructions may possibly involve multiple methods. As such, the measure of word count may inherently confound with the number of methods and steps. It was critically important to normalize word counts based on the number of methods mentioned in a tutorial and the number of steps mentioned in a method, so that the comparison of textual information has face validity among instructions covering different content. Because the normalized word counts neither positively nor negatively correlated with the tutorial age, it's considered unnecessary to normalize along the time.

To examine how the media modalities associate with normalized word counts, we used ANOVA and post-hoc Tukey HSD tests by using media modalities used in an article as the independent variable and normalized word count in the tutorial as the dependent variable. A main effect for media usage was found on normalized word count ($F(3, 1164) = 7.8322, p < 0.0001, \eta = 0.02$). Significant differences were found between rich media combinations (including at least one type of graphical representation, either static images or dynamic GIFs) and text-only baseline (all *p*s<.05). Adding images and animated GIFs, GIFs or images alone, to tutorials is significantly correlated with using more words in describing procedures than text-only baseline.

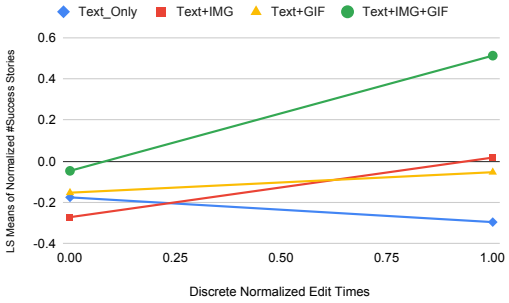


Figure 3: The interaction effect between media types and normalized edit times on the normalized number of successful stories

Analysis of Perception of Shared Knowledge by Learners

Finally, we considered the question how media modalities involved in How-To instructions correlate with the perception and learning from novice learners' perspective. We conducted regression analysis on the media choices and metrics that related to novice learners' perception and evaluation, including *helpfulness rating* and *number of success stories* of each How-To tutorial. Because articles were voted by learners after the instructional tutorials were encountered by different learners, these two metrics were accumulating along the time. As such, the number of edits made by editors were included in the regression model as a control approximating article age.

In order to generalize across different topical contents, the topic was included as a random effect into the model. Besides, the number of success stories was normalized along the time to make a fair comparison. Due to the fact that limited tutorials had a high number of successful stories, the long tail distribution of the successful stories was also log-normalized.

First, the model including media types, normalized edits and helpfulness rating was tested, and there was little collinearity effect in the model (VIF for all factors were smaller than 3.62). The linear regression reported a significant effect of media modalities on the helpfulness rating ($F = 8.5019, p < 0.0001, \eta^2 = 0.02$), which indicated that including images and GIFs into the tutorial significantly influenced viewers' perception of whether the How-To instructions were helpful. The post-hoc Tukey HSD test revealed that novice learners rated articles involving both GIFs and images as being more helpful than text-only counterparts. Including GIFs or images alone also results in significantly higher helpfulness ratings than text-only. However, the difference between including images only versus including GIFs only was not significant on this helpfulness rating.

Further, we explored how different combinations of modalities used in tutorials influence novice learners' successful acquisition of shared procedural knowledge through a linear regression analysis. A main effect was seen for media modalities, normalized edit times and their interaction effect on normalized number of success stories ($F = 5.4514, p < 0.0001, \eta^2 = 0.02$, $F = 21.825, p < 0.001, \eta^2 = 0.01$,

$F = 9.5863, p < 0.0001, \eta^2 = 0.02$). The prediction profiler indicated that less people had learned the procedural knowledge with text-only tutorials under more edits, while more novice learners successfully learned procedural knowledge with How-To instructions under more edits when involving media modalities other than text. To examine the statistical significance for the continuous variable of edit times, we discretized and normalized edit times according to the population median and re-ran the regression analysis for illustration purposes. Fig. 3 showed the increase of edit times significantly correlated with a drastic improvement of the number of successful stories when How-To tutorials included both static images and dynamic GIFs. However, the increase of edit times negatively correlated with the effectiveness of text-only tutorials.

Discussion

Results from our analyses support H1 and H2. The use of short dynamic graphical illustrations in the format of GIF shows the potential to positively transform procedural knowledge sharing. While our analyses are correlational in its nature, the results provide the necessary ground to further consider ways of using animating GIFs to support knowledge sharers and learners in procedural knowledge transfer in future.

H1 states that incorporating GIFs into instructions corresponds to more edits and more textual contents during the tutorial authoring process. In the analysis, we observed increased textual contents when GIFs were introduced to complement tutorial text, however, the edit effort (approximated using the number of edits) did not increase compared to the text-only baseline. The visual representation potentially helps editors better connect and ground their knowledge expositions with the procedural skills they performed and exercised in the real world, which may have facilitated collaborative editing among editors. Grounding with visualizations makes it possible for editors possessing different ideas about the procedure to see the commonalities and differences, allowing different parties to contribute to the tutorial. It's not surprising that introducing both static images and dynamic GIFs corresponded to even more verbalization. However, introducing two types of media did increase editing steps. The reason behind could be that editors need to decide where to put what when heterogeneity of different modalities exists, so that more edits are introduced to ensure content synchronization among combinations of modalities. Another implication is that introducing a mechanism to support collaborative mulimodal editing is necessary to enhance existing wiki-mediated collaborative writing platforms.

H2 is also supported by the analysis result. The observation indicated that richer media correlated to higher learners' evaluation and learning outcomes. When How-To tutorials include both static images and dynamic GIFs, novice learners have greater access to concrete information related to procedural behaviors and object states which approximates the benefits of face-to-face procedural knowledge transfer but in a scalable form.

In addition, the short GIF clips complementing tutorial text may deliver information beyond language barriers which may better assist non-native English-speaking learners. GIFs-incorporated How-To instructions also provide the utility to index specific steps in the procedure with visual representations, which leverage visual and verbal memories at the same time during task execution and help more viewers successfully acquire and perform the procedural task as a result.

Limitations and Future Work

Although we carefully controlled tutorial age, topic and quality when conducting the correlational observation by analyzing the archived data from wikiHow, the current analysis has limitations in making causal inferences. As such, comparative evaluation using experimental methods is necessary to compare how combinations of different modalities used in tutorials directly affect procedural knowledge sharing and transfer. In this case, finer-grained measures and observations will be possible for investigating the causal effects of GIFs and multimodalities in procedural knowledge sharing and consumption, and the conclusion can be potentially used to support the redesign of procedural knowledge sharing systems. In the future, we plan to conduct a qualitative interview study, so that the effects and mechanisms of multimodalities on procedural knowledge sharing can be cross validated by the users' perceptions and experiences.

Multilingualism potentially plays an important role in collaborative procedural knowledge sharing, where editors in different languages may collaborate differently. In this work, we investigated the effect of multimodal representations mainly on wikiHow tutorials in English, since viewers' perception and evaluation are not available for tutorials in other languages. To our best knowledge, non-English tutorials on the wikiHow platform are normally translated from the English version and the rating interaction is not provided for viewers. The research beyond English tutorials is open for future work.

Acknowledgments

We would like to sincerely thank all reviewers for their constructive feedback. We thank Jingxian Liao, Jingchao Fang and Zhouhao Wang for their valuable discussion and feedback. We also thank UC Davis GGCS for their support in this research.

References

Ackerman, M. S., J. Dachtera, V. Pipek, and V. Wulf (2013): 'Sharing knowledge and expertise: The CSCW view of knowledge management'. *Computer Supported Cooperative Work (CSCW)*, vol. 22, no. 4-6, pp. 531–573.

Bakhshi, S., D. A. Shamma, L. Kennedy, Y. Song, P. De Juan, and J. Kaye (2016): 'Fast, cheap, and good: Why animated GIFs engage us'. In: *Proceedings of the 2016 chi conference on human factors in computing systems*. pp. 575–586.

Bowers, J. M. (1991): 'The Janus Faces of Design: Some Critical Questions for CSCW'. In: J. M. Bowers and S. D. Benford (eds.): *Studies in Computer Supported Cooperative Work: Theory, Practice and Design*. Amsterdam, etc., pp. 333–350, North-Holland.

Brandes, U., P. Kenis, J. Lerner, and D. Van Raaij (2009): 'Network analysis of collaboration structure in Wikipedia'. In: *Proceedings of the 18th international conference on World wide web*. pp. 731–740.

Chi, P.-Y., S. Ahn, A. Ren, M. Dontcheva, W. Li, and B. Hartmann (2012): 'MixT: automatic generation of step-by-step mixed media tutorials'. In: *Proceedings of the 25th annual ACM symposium on User interface software and technology*. pp. 93–102.

Chirumalla, K., Y. Eriksson, P. Eriksson, et al. (2015): 'The influence of different media instructions on solving a procedural task'. In: *DS 80-11 Proceedings of the 20th International Conference on Engineering Design (ICED 15) Vol 11: Human Behaviour in Design, Design Education; Milan, Italy, 27-30.07. 15*. pp. 173–182.

Clark, H. H. and S. E. Brennan (1991): 'Grounding in communication.'.

Gergle, D., R. E. Kraut, and S. R. Fussell (2013): 'Using visual information for grounding and awareness in collaborative tasks'. *Human–Computer Interaction*, vol. 28, no. 1, pp. 1–39.

Gerson, E. M. and S. L. Star (1986): 'Analyzing due process in the workplace'. *ACM Transactions on Office Information Systems*, vol. 4, no. 3, pp. 257–270.

Hinds, P. J., M. Patterson, and J. Pfeffer (2001): 'Bothered by abstraction: The effect of expertise on knowledge transfer and subsequent novice performance.'. *Journal of applied psychology*, vol. 86, no. 6, pp. 1232.

Huang, D.-H. and W.-K. Chiou (2010): 'The effect of using visual information aids on learning performance during larger scale procedural task'. In: *3rd International Conference on Human System Interaction*. pp. 295–299.

Jiang, J. A., C. Fiesler, and J. R. Brubaker (2018): ''The Perfect One' Understanding Communication Practices and Challenges with Animated GIFs'. *Proceedings of the ACM on human-computer interaction*, vol. 2, no. CSCW, pp. 1–20.

Johansen, R. (1988): *Groupware. Computer Support for Business Teams*. New York and London: The Free Press.

Liu, J. and S. Ram (2011): 'Who does what: Collaboration patterns in the wikipedia and their impact on article quality'. *ACM Transactions on Management Information Systems (TMIS)*, vol. 2, no. 2, pp. 1–23.

Marsh, E. E. and M. D. White (2003): 'A taxonomy of relationships between images and text'. *Journal of Documentation*.

Miltner, K. M. and T. Highfield (2017): 'Never gonna GIF you up: Analyzing the cultural significance of the animated GIF'. *Social Media+ Society*, vol. 3, no. 3, pp. 2056305117725223.

Navarrete, T. and E. Villaespesa (2020): 'Image-based information: paintings in Wikipedia'. *Journal of Documentation*.

Palmiter, S., J. Elkerton, and P. Baggett (1991): 'Animated demonstrations vs written instructions for learning procedural tasks: a preliminary investigation'. *International Journal of Man-Machine Studies*, vol. 34, no. 5, pp. 687–701.

Salaberry, M. R. (2018): 'Declarative versus procedural knowledge'. *The TESOL Encyclopedia of English language teaching*, pp. 1–7.

Shaw, A. and E. Hargittai (2018): 'The pipeline of online participation inequalities: The case of Wikipedia editing'. *Journal of communication*, vol. 68, no. 1, pp. 143–168.

Tolins, J. and P. Samermit (2016): 'GIFs as embodied enactments in text-mediated conversation'. *Research on Language and Social Interaction*, vol. 49, no. 2, pp. 75–91.

Torrey, C., D. W. McDonald, B. N. Schilit, and S. Bly (2007): 'How-To pages: Informal systems of expertise sharing'. In: *ECSCW 2007*. Springer, pp. 391–410.

Viegas, F. B. (2007): 'The visual side of wikipedia'. In: *2007 40th Annual Hawaii International Conference on System Sciences (HICSS'07)*. pp. 85–85.

Volker, M. S. A. V. P. et al. (2003): *Sharing expertise: Beyond knowledge management*. MIT press.

Wilkinson, D. M. and B. A. Huberman (2007): 'Cooperation and quality in wikipedia'. In: *Proceedings of the 2007 international symposium on Wikis*. pp. 157–164.

Yang, C.-L. and H.-C. Wang (2019): 'Understanding How Social Prompts Influence Expert's Sharing of How-to Knowledge'. In: *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. pp. 433–437.

Yang, J.-T. (2007): 'Knowledge sharing: Investigating appropriate leadership roles and collaborative culture'. *Tourism management*, vol. 28, no. 2, pp. 530–543.

Bittenbinder, S., Pinatti de Carvalho, A. F., Krapp, E., Müller, C., Wulf, V. (2021): Planning for Inclusive Design Workshops: Fostering Collaboration between People with and without Visual Impairment. In: Proceedings of the 19th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591), DOI: 10.18420/ecscw2021_ep27

Planning for Inclusive Design Workshops: Fostering Collaboration between People with and without Visual Impairment

Sven Bittenbinder[1], Aparecido Fabiano Pinatti de Carvalho[2], Eva Krapp[1, 2], Claudia Müller[1], Volker Wulf[2]

[1] Institute of Information Systems, esp. IT for the Ageing Society, University of Siegen, Germany
[2] Institute of Information Systems and New Media, University of Siegen, Germany

{sven.bittenbinder, fabiano.pinatti, claudia.mueller, volker.wulf}@uni-siegen.de, eva.krapp@student.uni-siegen.de

Abstract. Carrying out successful design workshops can be a challenging task. This can turn even more difficult, if one attempts to engage in more inclusive design workshops, where a broad range of user profiles are covered. If some of these profiles refer to people with impairments, things can get even more complicated. Furthermore, there are also associated challenges when trying to carry out something that is usually implemented as a face-to-face activity in an online format. This exploratory paper introduces a discussion on a few lessons learned from organising design workshops including both people with and without visual impairments. It also outlines our response to the situation created by the COVID-19 pandemic, which prevented us to engage in face-to-face design workshops.

Based on feedback received from participants of a first in-person design workshop organised within one of our projects and on informal interviews carried out mainly over the phone to discuss ways to enhance the collaboration between people with and without visual impairments during such activities, we go on to introduce some relevant aspects that should be taken into consideration when planning inclusive design workshops. This is a preliminary contribution, meant to raise discussions on technology-mediated inclusive participatory design initiatives to further inform the development of a solid methodological contribution to CSCW.

Introduction

The value of a Participatory Design (PD) approaches for the conception and elaboration of interactive systems has been acknowledged since long ago within the HCI and CSCW communities (Bødker, 1996; Bjögvinsson et al., 2012; DiSalvo et al., 2013). Currently, more and more HCI and CSCW researchers and practitioners draw on such approaches, as they go on to engage in user-centred design (Muller, 2002). Such approaches are even more relevant when designing for people with impairments, as there is a growing understanding that one should not design for people with a particular impairment if one does not share the same impairment or, at least, listen to those who experience it in their everyday lives **(**Bennett, 2018; Kane et al., 2014**)**. The complexity of this equation increases when a system targets not only people with but also those without impairments and designers attempt to foster collaboration between the two of them during the design process.

This is what we have been experiencing in the project iDES^{kmu}. The project, which is funded by the German Federal Ministry of Labour and Social Affairs, sets out to investigate issues of accessibility in document and enterprise content management systems (DMS/ECMS) often used within small and medium enterprises (SMEs). The project goals concern, among other things, the understanding of the extent to what people with visual impairments would be able to use such systems in a work context. The project also aims at sensitising the general population about the relevance of accessible workplaces. Accessible software would be a relevant part of such workplaces, as demonstrated by findings from the literature **(**Branham & Kane, 2015).

The project hence addresses three different target groups: *companies* using or developing DMS and ECMS; *users* of such systems, both in an SME context or not; and *software development actors* or, in other words, people who contribute for the design and development of software systems – e.g., interface and user experience designers, usability and accessibility professionals and developers in general (de Carvalho et al., 2020). These target groups are naturally not mutually exclusive. For example, freelancers working with software development would be a representative of both companies and the software development actors. By the same token, accessibility evaluators, who also work with the types of software that

they evaluate, would be a representative of both the users and the software development actor target groups.

In order to reach part of its goals, the project proposes to design and develop an application that would allow people to evaluate the most relevant accessibility aspects of a particular system. We refer to this application as the Testing Suite (TS). Members of all three target groups mentioned above should be able to use the TS, independent of their level of expertise with approaches to testing software accessibility or the guidelines involved in it. By providing the community with such a tool, the project aims at providing people with the possibility to carry out elementary accessibility checks as they go on to: (1) engage in developing a particular system; (2) consider to buy a particular software; (3) make a case to change a particular software application in the workplace for one with better accessibility; among other things. Such a tool would thus contribute towards accessible software development and decision making processes in regard to buying or replacing a piece of software.

As part of the PD activities of the project, design workshops (DWs) have been planned to build the referred application together with the different stakeholders of the system. Despite our experience with DWs, engaging people with and without visual impairments in DW activities have proven somehow challenging. A survey of the literature have revealed a visible gap of research on how to foster collaboration between people with and without visual impairments in these activities. Even if some work on how to engage visually impaired people in design activities can be found **(**Magnusson et al.; 2018; Bischof et al., 2016**)**, there is a lack of work on the interaction between participants with and without such impairments. This exploratory paper drafts the first lines of an answer to a related question. We therefore set out to provide preliminary results to answer the question: *how can we foster collaboration between people with and without visual impairments during DW activities?* It is not our intention here to provide a definite or final answer to this **research question**, but rather to engage with the community in exploring some findings that shed light on this and define future directions in the development of a solid contribution to the field.

Due to the COVID-19 pandemic, we have been faced with the need to carry out online DWs. This added a layer of complexity in the planning for our DWs, as will become clear across the Findings section. Our contribution, therefore, is not constrained to face-to-face DWs, but also online DWs, which until before the outbreak of the COVID-19 pandemic was quite uncommon **(**Singh, 2020; Martin et al., 2020). Furthermore, as already mentioned our contribution addresses a very relevant issue which, to-date, has not been satisfactorily addressed in our community: the collaboration between people with and without visual impairments in PD activities. With this study, we set out to investigate how we can carry out PD workshops in which we bring the full potential of all the involved parts. This means that, whilst we would like to provide visually impaired people with the opportunity

to fully participate in the activity, we are also keen to prevent constraining sighted people in their participation, as for example, by avoiding the use of visual artefacts.

The remainder of this contribution is organised as follows: section 2 (Context) gives an overview of the complex design research process involved in conceptualising the referred TS, making reference to the methods used for each of the relevant phases; section 3 (Methodology) provides information about to the methodological approach that resulted in the findings related to our research question; section 4 (Results) introduces the analysis of the preliminary findings of our investigation; section 5 (Discussion) carries out a short discussion of the presented findings; finally, section 6 (Conclusion) lays out some concluding remarks and our plans to pursue this investigation.

Context

Taking into account the relevance of a user-centred and practice-based design approach for the conception of useful and usable applications (Rohde et al., 2016), we draw on the Design Case Study (DCS) framework (Wulf et al., 2015) for the overall design of our TS. The framework is organised in *three distinct phases*, which can coexist in some moments of the design and development process. The first phase, known as the *pre-study,* concerns the understanding of people's work contexts and practices. For that, qualitative or mixed-methods studies predicated on methods like *in-dept interviews* (Hermanowicz, 2002), *participant observation* (McKechnie, 2008) and *cultural probes* (Gaver et al., 1999) are carried out. Most often, an ethnographic approach is used during it, but it is not uncommon for this phase to be implemented as an interview study. At the end of it, design opportunities outlining the design space are identified. The *design* phase is predicated on methods like sketching and prototyping, as well as assorted usability evaluation techniques, as for example, Heuristic Evaluation (Molich and Nielsen, 1990) and Cooperative Evaluation (Monk et al., 1993). Most often, a PD approach is used, based on a series of DWs with representative people from the target group(s). Last but not least, the *appropriation* phase refers to the deployment of the artefacts generated during the design phase to naturalistic environments, and the study of how the usage of such artefacts will (or will not) change practices.

For our own purposes, we have used in-depth interviews and participant observations as the main data collection methods for the pre-study. The interview study included members of the three target groups previously mentioned (users, companies and software development actors) and focused on understanding the participants' awareness and knowledge about software accessibility and accessibility testing. During the interviews, participants also talked about the relevance of a tool such as the TS and how such an application should look like. The observations focused on practices of accessibility professionals in terms of carrying out accessibility tests. These observations have generated further

information about the features that an application like the TS should include, in order to support people in carrying out elementary accessibility tests. The collected data has undergone a *thematic analysis* (TA) according to Braun and Clarke's approach (Braun and Clarke, 2012) and generated a series of themes concerning design implications and requirements for the TS.

We are currently undergoing the design phase of our project. This phase has been predicated on a series of DWs to discuss and elaborate on the results of the pre-study with the participants. Furthermore, the DWs have been used to envisage and conceptualise the user interfaces and interactive mechanisms for the TS. Among the approaches that we have been using for the DWs are: *brainstorming*, *scenario-based design* (Carroll, 2000) and *low-fidelity prototyping*. These are traditional methods in PD initiatives, as widely acknowledged in the literature (Muller 2002).

The study originating the findings of this contribution emerged from our experiences with organising and running our first DW. This DW featured 8 participants, some of whom have also participated in our pre-study, as seen in Table I.

Table I. Participants of the first on-site DW

Participant #	Visual acuity[1]	Access. Expertise	Pre-study participant	Informal interview after 1. DW	Target Group
P1	legally blind		x	x	User
P2	legally blind				User
P3	fully blind	x		x	User
P4	sighted	x		x	Soft. dev. / Company
P5	sighted				Soft. dev.
P6	sighted	x			Soft. dev.
P7	sighted	x	x		User
P8	sighted	x	x		Soft. dev.

The decision to include participants from the pre-study as well as new participants was deliberate, as the literature suggests that this can bring an interesting dynamic to DWs: while participants from the pre-study could resonate with some of the data presented, new participants could either confirm or challenge it (Sharp et al., 2006). This was exactly the purpose of the brainstorming session that followed the introduction of the pre-study results.

1 Visual acuity refers to the extent to what a person can clearly see. Governments usually refer to a common acuity scales, in order to decide whether someone is entitled to receive some benefits from programmes sponsored by them. For instance, the US government use a scale that includes the categories partially sighted, low vision, legally blind *and* totally blind. How each of these categories are defined usually varies depending on the country. In Germany, for example, legally blind refer to people whose visual acuity is lower than 1/50 (Rohrschneider, 2018).

We tried to engage both visually impaired and sighted representatives from all different target groups. Unfortunately, we were not able to recruit any visually impaired person working in software development. Our experience suggests that there is lack of visually impaired software developers. In addition to that, we were only able to recruit one representative of the company target group, who turned out to be also a developer: he was a freelancer working with accessible software development.

We have planned this workshop as a series of activities focusing on brainstorming and group work. Table II portrays the agenda for our DW and, consequentially, the activities carried out.

Table II. Agenda for the first iDESkmu DW

09:15	**Welcome**
09:30	Agenda overview
09:40	Introduction round
09:55	Ice-breaker game
10:15	Brainstorming session on results from interviews and observations carried on in the project
11:00	**Pause 1**
11:15	*Parallel Group Session 1:* Prioritising and expanding design requirements for the Test Suite
12:00	*Presentation Round 1:* Presentation of results of group session 1
12:30	**Lunch**
13:00	*Parallel Group Session 2:* Selection of best ideas presented in presentation round 1
13:45	*Presentation Round 2:* Presentation of results of the group session 2
14:15	**Pause 2**
14:30	Final integration of results and prioritisation of requirements
15:15	Wrap-up
15:30	**End**

After welcoming the participants to our premises, we have started with a short overview of the agenda and an explanation of the dynamics of the workshop activities. In this moment, we have explained how participants would be working together and which sorts of artefacts they would have to produced. They have also been introduced to the materials that they could use, as for example, flipcharts, post-it notes, colour pens, etc. We consciously offered participants the opportunity to work and generate visual artefacts, but we have of course diligently worked to sensitise participants about the relevance that all of them – independently if with or without impairment – participated in all of the activities in their full capacity. Put differently, we were very keen that the visually impaired participants would be able to contribute to the activity, at the same time that we did not want that the sighted people would be constraint in what they could produce and use in the workshop. Therefore, we made very clear from the beginning that, while visual artefacts could be produced and used, they would have to be expressed in other formats as well, so

that the visually impaired people could understand, follow and contribute to the activity in question. Nevertheless, we have not established any prior mechanism for the interaction between people with and without impairments, which turned out to be one of the barriers for full participation in the workshop, as discussed ahead in the paper.

The initial session was followed by a short introduction round where everybody had the opportunity to introduce themselves. As next, we engaged in an ice-breaker game where participants had to say a potential fun fact about themselves and the other participants would have to try to guess if that was true of false. This turned out to be important for the collaborative work that followed.

The first work session was a brainstorm. The themes regarding the design features of the TS identified in the analysis of the pre-study data was introduced by one of the organisers and used as the discussion basis of the activity. Overall, participants discussed among themselves the extent to what the presented requirement corresponded to what they would envisage as an accessibility TS. After the initial brainstorm, participants were split in 2 groups and were asked to discuss which of the features presented would be relevant for the TS. They were also asked to add any new feature that they thought was missing. This was done, so that participants who were not part of the pre-study also had the opportunity to contribute to the requirements elicitation process of the TS.

In order to guarantee diversity in the groups, we purposively distributed users and software development actors evenly across the two groups. Consequentially, all groups had at least one visually impaired and one sighted person. This was from our perspective very important to bring different points of view to the discussion. Furthermore, having a visually impaired person participating in the discussions could enhance the likelihood that accessibility requirements would emerge. It is sensible to think, we argue, that these participants would share their experiences on how the TS should be designed, so that they could also use it.

After the presentation of the results of the first activity, members of the two groups were mixed, resulting in two new group configurations. We did that deliberately to create new dynamics in the discussions. In the second activity, group members were asked to create an integrated list of features, based on the results presented by the two groups, and asked to rank it according to the relevance. By the end of the activity, the integrated lists were compared and a final list of ranked features were elaborated. This integrated list was to be used in a second DW, where participants were meant to engage in scenario-based design and low-fidelity prototyping.

Although the second DW was also planned to be a face-to-face activity, we were faced with the second wave of the COVID-19 pandemic, and due to the social distancing restrictions, we were no longer allowed to meet personally, which demanded changes in the plans. The articulation work in preparing the second DW is one of the focus of this particular contribution. In addition to that, the feedback

during the debriefing session run immediately after the first DW to discuss the extent to what the activities were easy or difficult to achieve, we could notice some challenges regarding the collaboration between people with and without visual impairment. This brought us to the research question whose answer we explore in this contribution.

Methodology

In order to find an answer for our research question – i.e., *how can we foster collaboration between people with and without visual impairments during DW activities?* – we have adopted a reflective approach taking into consideration both the feedback collected during the debriefing session of the first DW carried out for the conception of the TS as well as a series of informal interviews carried out over the telephone, as we were planning our second DW.

As mentioned above, the debriefing session took place immediately after the first DW and lasted for about half an hour. It has been carried out as a short *focus group* and concentrated in the participants views on what they liked about the DW they had just participated in and what they thought it should be improved for future DWs. The session has been audio recorded and transcribed for subsequent analysis. Fieldnotes have also been taken by the three researchers who were facilitating the DW activities.

In terms of the informal interviews, as soon as we noticed that it would not be possible to carry out the next DW, we have articulated with participants how an online workshop could be made viable. Our main concerns were to make sure that the participants would be able to collaborate with each other during the workshop, independent of the sight condition.

Since participants were fully employed and already granting us valuable time for the DWs, we did not want to burden them with a long in-dept semi-structure interview. Furthermore, since we were not allowed to meet face-to-face, we have to carry out the interview remotely. Since they were meant to be very short interviews, we decided that using the phone, instead of a video conference system would be more appropriate. In addition to that, especially for our visually impaired participants, we were not sure about their experience with video-conference systems and which of them would be more appropriate for the interaction. Answers to the informal interviews have been recorded as field notes and served as information for the preparation of the second DW.

In terms of data analysis, we used a group reflection approach, in which the three first authors of this contribution have gone recurrently through the collected data together – both during the focus group as well as during the informal interviews – and sought for patterns in the answers, which have been also confronted with the knowledge available in the relevant literature. We present the preliminary results of this deep reflective approach in the following section.

Although some may have some reservations about the data analysis approach employed for this contribution, we would like to remind that it is not our intention to provide definite answers to our research question through this contribution. On the contrary, we would share preliminary findings with the community and further explore issues which we should pay attention to as we move on with our analysis. It is our intention to submit our data to a systematic TA, using the same approach as the one used for the pre-study data, so that we can provide the community with deeper insights on the themes herein introduced.

Results

The analysis carried out on our focus groups data and the informal interviews suggests that there are many aspects that should be considered in order to foster the collaboration between people with and without visual impairment during DW activities. The first of it refers to supporting different media artefacts, which allows participants from both groups to make the most of the senses available to them.

A general aim of DWs is to foster mutual learning among participants stemming from different domains and to reduce the "symmetry of ignorance" (Fischer, 2000). Tools which foster knowledge exchange and creativity used in such workshops predominantly address visual elements – e.g., paper and cardboards and whiteboards (Muller, 2002). Using these visual elements ideas for solutions are generated together with the participants, in that all the information developed is made available to all participants equally.

As explained before, we found it relevant not to constrain the sighted people, by prohibiting the production and use of visual artefacts. Nevertheless, our findings highlighted that a challenge to engage people with and without visual impairment in such activities is to effectively distribute the relevant and mostly visual information to the visually impaired participants. One of the critiques that we received from our participants during the debriefing session of the first in-person workshops referred to the accessibility of information provided during it or created by the sighted participants during the activities:

> *We are talking about accessibility and I had the feeling that I have to learn by heart what participants say and what comes out of the individual groups, because there was nothing barrier-free where I could get access to it again. Except that at some point I took out my computer and wrote down all the past information in no time at all. Who said when and what so that I could even see where we are? On this point, I would like us to become more accessible there too. For those who see nothing or can write on a piece of paper. (P3)*

This draw our attention to the fact that the predominant usage of tools for visualisation in a co-present setting is hardly accessible for visually impaired

participants. Despite the fact that no activity of our first DW required any sort of sketching or visual mapping, sighted participant indeed used note taking during the discussion sessions in physical notepads, which has been used to report on the discussions in the group. Although visually impaired people could also have laptops for that activity, we have not asked them to bring their laptops or mentioned that they would possibly want to take note of discussions. We took that for granted and, as a result, we disabled the visually impaired people to fully participate of the experience. Furthermore, we used slide presentations to introduce our findings, which have not been previously shared with the participants, as we did not want them to come with pre-determined views on the issues that would be discussed. Nevertheless, this proved to be an erroneous decision.

Reflecting upon these findings, we get to the conclusion that verbal description of the information is indispensable for the explanation of the ideas that have been noted as bullet points in a block of paper or in a set of slides, but this is not sufficient for such a wealth of information and the impossibility to refer back to such notes and read them up can disadvantage blind and visually impaired participants. This finding has also been confirmed by the other visually impaired participants during the focus group. It suggests that there is a need for a common, electronic and accessible document that can be used by all participants. Any visual information must be possible to be translated to other type of outputs. It is sensible to think that artefacts generated during particular DW session would be available on all participants own computers so that they can individually explore it. These artefacts must be prepared in a way that, if a participant uses a particular assistive software, such as screen readers or magnification software, they would be able to use it without any problems (Coombs, 2010).

As mentioned before, these findings are not limited to the accessibility of the materials created during the DW, but also to those created by the organisers, such as agenda, time and task planning. All of these materials should be provided beforehand in an accessible format, in order to allow for a successful collaboration during the DW activities. Furthermore, templates for particular artefacts to be generated during such activities should be provided. For instance, for our second DW, we were planning that participants would engage in scenario-based design. One of the organisation aspects that we had in our minds were to provide templates for the scenarios that would be written during the DW activities.

Regarding possible formats of the most accessible documents, Coombs (2010) refers to the versatility of Office products (e.g. MICROSOFT OFFICE, LIBREOFFICE, etc.). In an informal conversation after the de-briefing, however, P3 reiterated that the mere possibility of using a shared, accessible online document on one's own computer does not mean that a detailed verbal description of graphics and other information during the meeting is no longer required, whether on-site or online.

Another aspect to be taken into consideration are ways to make it possible for visually impaired people to engage in low-fidelity prototyping. As previously

mentioned, for our second DW, participants will be required to engage in low-fidelity prototyping. We were not completely sure how we should approach the activity, so that people with and without visual impairment could cooperate successfully. Our main question was how blind participants would be able to contribute to the elaboration of a (visual) prototype and to "read" it beyond the verbal description by means of their own aids. Race et al. (2020) state that the most common methods for “making visually rendered information visible” are textual descriptions or tactile graphics – i.e., graphics to touch. On the other hand, by discussing the matter with our participants, we were suggested to use a spreadsheet to divide the screen into quadrants. Navigation through the individual cells into the spreadsheet editor would be possible using the keyboard and the screen reader in their laptops would read the content aloud. Participants could then textually describe what would be present in each quadrant of the screen.

These findings suggest that one would need a platform to centralise all of those artefacts. One of the possibilities would be to use something like GOOGLE DRIVE or MICROSOFT ONEDRIVE which integrate different editing applications, spanning text and spreadsheet editors to presentation editors. In addition to that, the demands we have observed, which suggest that DW based on the collaboration between people with and without visual impairment should be predicated upon the use of digital technology and the generation of digital artefacts, can be seen as a step towards the digitalisation of DW. This, from our perspective, would make it easier to make these DW totally online, since a whole infrastructure for the generation and sharing from the aforementioned artefacts should already been generated. The last step towards a complete online DW workshop would be the use of a proper video conferencing tool. This is particular interesting in a situation as the one we faced due to the second wave of the COVID-19 pandemic.

Among the available tools, participants have mentioned ZOOM as one of the platforms available in the market, which is also accessible. For instance, P3 and P4, both experts in accessible software, mentioned it as a good solution for video conference involving people with visual impairment, in conformity with findings presented by Hersh et al. (2020).

Our findings therefore suggest, that for fostering collaboration between people with and without visual impairment, we need an infrastructure for cooperative work. Figure 1 represents such an infrastructure, taking account the performance of online DW featuring people with and without visual impairment. In the case of face-to-face DWs, the infrastructure would be very similar. The only difference would be the elimination of the video conferencing tool to mediate the communication between participants of the workshop.

It is worth pointing out that participants have mentioned both during the focus group as during the informal interviews that, one of challenges organising online workshops would be to keep people engaged across long hours, as is the case of its face-to-face counterpart. Past and current research have demonstrated that long

video sessions are perceived as very exhausting and tiring for the participants (Wiederhold, 2020). This is something to be aware of when transferring a classic face-to-face workshop to the online world. A potential solution for this would be splitting the DW activities in several self-contained activities, which could be accomplished within one to two hours. Our findings suggested that this would actually enhance participation in DW. Participants recurrently mentioned that, despite their interest in participating in such events, the fact that they worked full time would prevent them to do so in a more frequent basis. It would be easier to coordinate short session between participants and to get them involved during the PD activities of the project. This is another relevant finding that should be taken into consideration when planning inclusive and sustainable DW.

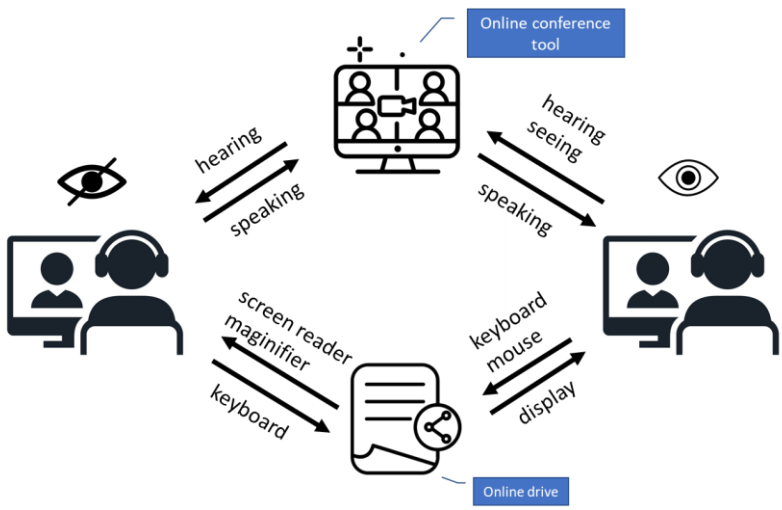


Figure 1 Set-up for Inclusive Online DW feature people with and without visual impairment

Discussion

As mentioned before, it is well accepted within the HCI and CSCW fields, that involving users in the design process of new solutions is a very important for the conception of useful and usable solutions. The involvement of people with impairments in the process of developing solutions for them has been considered even more necessary (Bennett, 2018; Kane et al., 2014). However, there is still a lack of literature and knowledge on how people with and without impairments can be effectively included in PD activities due to their different needs.

The shift in methods of collaboration from the analogue to the digital world, due to the Covid-19 pandemic, offers the opportunity to try out new ways of collaboration over distances and could be the ideal opportunity to also find new tools and methods (practices) for collaboration between people with and without

disabilities to develop and test in context of PD (Singh, 2020; Martin et al. 2020). The findings that we present suggest that this is actually a necessary change in paradigm, if we would like to foster collaboration between people with and without impairment in DW activities.

As introduced in the Results section, the provision of a second communication channel during a DW – be it online or co-located – can fulfil the formal requirements of providing blind and sighted participants in DWs with information of the same quality, even if the type of consumption is different. This communication channel should be based on the provision of accessible documents, which can be handled by different assistive technology, which participants with visual impairment may need to use, in order to make sense of the artefacts being conceived and contribute towards their conception.

Therefore, as answer to our research question of how cooperation between people with and without disabilities in DW activities can be promoted, we can potentially say that providing an infrastructure for the elaboration and sharing of digital artefacts, independent if in a textual or in a graphic format, is key.

Nevertheless, other factors can also play a role for success in practice, such as the ease with which the online conference tool and the online drive can be appropriated, i.e., their user friendliness, accessibility and the user experience that they can offer. It is therefore entirely possible that the tools used have an impact on the success of planned activities.

In addition to the mostly used possibilities of making material readable for blind people either by textual description or tactile graphics (Race et al., 2020), our results have shown another possibility of making visual information of a prototype accessible for blind and visually impaired people, namely in the form of screen reader-readable tables. Furthermore, it should be taken into consideration that many blind people are not blind by birth, so they are likely to know some software applications, may have used it in the past, and could imagine how to interact with software in an effective way to test software for accessibility. And much more, this kind of prototyping in digital form could open up completely new possibilities for prototypes other than software products.

Conclusion and Future Work

This contribution advances the state of the art by introducing a discussion of how collaboration between people with and without visual impairments can be fostered during DW activities. The findings we presented provide strong indicators of the need for adjustments when design for inclusive DW between people with and without visual impairments. We argue that the presented findings are of great value for the planning of DW which allow for successful collaboration between those actors.

As future work, we propose to subject the results presented in this contribution to scrutiny, by planning and carrying out a DW including people with and without visual impairment, supported by an infrastructure as the one introduced in our Results sections. The findings from this future initiative will allow us to assess the extent to what the proposed approach would effectively work for fostering collaboration between its participants and which other aspects must be taken into consideration to successfully achieving this goal.

References

Bennett, C. L. (2018) 'A toolkit for facilitating accessible design with blind people', *ACM SIGACCESS Accessibility and Computing*, no. 120, pp. 16–19. doi: 10.1145/3178412.3178415.

Bischof, A. *et al.* (2016) 'Exploring the Playfulness of Tools for Co-Designing Smart Connected Devices: A Case Study with Blind and Visually Impaired Students', in *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*. New York, NY, USA: Association for Computing Machinery (CHI PLAY Companion '16), pp. 93–99. doi: 10.1145/2968120.2987728.

Björgvinsson, E., Ehn, P. and Hillgren, P. (2012) 'Design things and design thinking: contemporary participatory design challenges', *Design Issues*, vol. 28, no. 3, pp. 101–116.

Branham, S. M. and Kane, S. K. (2015) 'The Invisible Work of Accessibility: How Blind Employees Manage Accessibility in Mixed-Ability Workplaces', in *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility - ASSETS '15. the 17th International ACM SIGACCESS Conference*, Lisbon, Portugal: ACM Press, pp. 163–171. doi: 10.1145/2700648.2809864.

Braun, V. and Clarke, V. (2012) 'Thematic analysis.', in Cooper, H. et al. (eds) *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* Washington: American Psychological Association, pp. 57–71. doi: 10.1037/13620-004.

Carroll, J. M. (2000) 'Five reasons for scenario-based design', *Interacting with Computers*, vol. 13, no. 1, pp. 43–60. doi: 10.1016/S0953-5438(00)00023-0.

Coombs, N. (2010) *Making Online Teaching Accessible: Inclusive Course Design for Students with Disabilities*. Wiley (Jossey-Bass Guides to Online Teaching and Learning).

DiSalvo, C., Clement, A. and Pipek, V. (2013) 'Participatory Design For, With, and By Communities', in Simonsen, J. and Robertson, T. (eds), International Handbook of Participatory Design, Oxford: Routledge, pp. 182–209.

Fischer, G. (2000) 'Symmetry of Ignorance, Social Creativity, and Meta-Design', *Know.-Based Syst.*, vol. 13, no. 7–8, pp. 527–537. doi: 10.1016/S0950-7051(00)00065-4.

Hermanowicz, J. C. (2002) 'The Great Interview: 25 Strategies for Studying People in Bed', *Qualitative Sociology*, 25(4), pp. 479–499.

Hersh, M., Leprini, B. and Buzzi, M. (2020) 'Accessibility Evaluation of Video Conferencing Tools to Support Disabled People in Distance Teaching, Meetings and other Activities', in *ICCHPOpen Access Compendium Future Perspectives of AT, eAccessibility and eInclusion. 17th International Conference on Computers Helping People with Special Needs*, Online, pp. 133–140.

Gaver, B., Dunne, T. and Pacenti, E. (1999) 'Design: Cultural Probes', *Interactions*, vol. 6, no. 1, pp. 21–29. doi: 10.1145/291224.291235.

Kane, S. K. *et al.* (2014) 'Collaboratively designing assistive technology', *Interactions*, vol. 21, no. 2, pp. 78–81. doi: 10.1145/2566462.

Magnusson, C., Hedvall, P.-O. and Caltenco, H. (2018) 'Co-designing together with Persons with Visual Impairments', in Pissaloux, E. and Velazquez, R. (eds) *Mobility of Visually Impaired People*. Cham: Springer International Publishing, pp. 411–434. doi: 10.1007/978-3-319-54446-5_14.

Martin, J., Loke, L. and Grace, K. (2020) 'Challenges facing movement research in the time of Covid-19: Issues in redesigning workshops for remote participation and data collection', in *32nd Australian Conference on Human-Computer Interaction*. New York, NY, USA: Association for Computing Machinery (OzCHI '20), pp. 712–716. doi: 10.1145/3441000.3441055.

McKechnie, L. E. F. (2008) 'Participant Observation', in Given, L. M. (ed.) *The SAGE Encyclopedia of Qualitative Research Methods*. Thousand Oaks: SAGE Publications, Inc., pp. 598–599.

Molich, R. and Nielsen, J. (1990) 'Improving a Human-Computer Dialogue', *Commun. ACM*, vol. 33, no. 3, pp. 338–348. doi: 10.1145/77481.77486.

Monk, A. *et al.* (1993) 'Cooperative Evaluation: A Run-time Guide', in *Improving your Human-Computer Interface: A practical Technique*. New York: Prentice-Hall.

Muller, M. J. (2002) 'Participatory Design: The Third Space in HCI', in *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*. USA: L. Erlbaum Associates Inc., pp. 1051–1068.

Pagon, R. A. (1988) 'Retinitis pigmentosa', *Survey of Ophthalmology*, vol. 33, no. 3, pp. 137–177. doi: https://doi.org/10.1016/0039-6257(88)90085-9.

Pinatti de Carvalho, A. F. *et al.* (2020) 'Fostering Accessibility at the Workplace through Community-based Participatory Research', *Proceedings of 18th European Conference on Computer-Supported Cooperative Work*. doi: 10.18420/ecscw2020_ws07.

Race, L. *et al.* (2020) 'Designing Educational Materials for a Blind Arduino Workshop', in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20: CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, pp. 1–7. doi: 10.1145/3334480.3383055.

Rohrschneider, K. (2018) 'Blindheit in Deutschland im 20. Jahrhundert', in *Blindheit in der Gesellschaft: Historischer Wandel und interdisziplinäre Zugänge*. 6th edn. Campus Verlag, p. 97.

Sharp, H., Rogers, Y. and Preece, J. (2006) Interaction Design: Beyond Human-Computer Interaction, 2nd ed. West Sussex, John Wiley & Sons.

Singh, V. (2020) 'Workshops are now required to be conducted remotely: is this a bad thing?', *Interactions*, vol. 27, no. 4, pp. 52–54. doi: 10.1145/3406102.

Wiederhold, B. K. (2020) 'Connecting Through Technology During the Coronavirus Disease 2019 Pandemic: Avoiding "Zoom Fatigue"', *Cyberpsychology, Behavior, and Social Networking*, vol. 23, no. 7, pp. 437–438. doi: 10.1089/cyber.2020.29188.bkw.

Joni Salminen, Ahmed Mohamed Kamel, Soon-gyo Jung, and Bernard J. Jansen (2021): The Problem of Majority Voting in Crowdsourcing with Binary Classes In: Proceedings of the 19th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591), DOI: 10.18420/ecscw2021_n12

The Problem of Majority Voting in Crowdsourcing with Binary Classes

Joni Salminen[1,2], Ahmed Mohamed Kamel[3], Soon-gyo Jung[1], Bernard J. Jansen[1]

[1]Qatar Computing Research Institute, Hamad Bin Khalifa University, [2]University of Turku, [3]University of Cairo
{jsalminen,sjung,bjansen}@hbku.edu.qa, ahmedm.kamel@pharma.cu.edu.eg

Abstract. When there are two classes, a majority vote can always be obtained with three labelers. Researchers can utilize this property to obtain a false sense of confidence in their ground truth labels. We demonstrate such a case with 3000 crowdsourced labels for an online hate dataset. Evaluating with percentage agreement, Gwet's AC1, and Krippendorff's alpha, results show that using more raters teases out the hidden nuances in raters' preferences. We show that full agreement among the raters monotonically decreases from three raters (28.4%) to nine raters (19.5%). Ten raters have a higher agreement than any other number of raters, which supports the idea of increasing the number of raters for subjective labeling tasks. Nevertheless, while beneficial, increasing the number of raters cannot be considered as a fundamental solution to the issue of agreement in subjective crowdsourcing tasks, as even with ten raters, there is a non-negligible number of ties (4.11%). We suggest having a small sample of the data labeled by five or more raters to evaluate the stability of agreement among the raters.

Introduction

Our argument focuses on the development of training sets for online hate detection (classification, scoring) models that are used in various computing systems. We argue that using binary classification with three raters can hide underlying disagreement among crowd raters. We suggest ways to tackle this issue.

Subjectivity in crowdsourced ratings is well-known (Alonso, 2011, 2015; Alonso et al., 2013, 2015; Alonso & Mizzaro, 2012; Aroyo et al., 2019; Salminen, Almerekhi, Dey, et al., 2018). Fundamentally, subjectivity means that individuals rate items differently based on personal beliefs, attitudes, worldviews, cultures, demographics, and other factors affecting their judgment (Alonso, 2015). Despite this, researchers dealing with subjective rating tasks, such as online hate/toxicity annotation, still use crowdsourced labels to construct training sets (Almerekhi et al., 2019, 2020; Davidson et al., 2019, 2017; Fortuna, 2017; Vidgen & Derczynski, 2020) for machine learning (ML). Crowdsourcing, in general, refers to using an anonymous pool of users to carry out human intelligence tasks (HITs) (Kittur, Chi, et al., 2009; Kittur et al., 2008, 2013; Kittur, Lee, et al., 2009; Yu et al., 2016).

In this research, we investigate the particular case of majority voting in a binary labeling task when using crowdsourced ratings. Binary labeling refers to a task where the raters have two options (e.g., yes/no, positive/negative). Majority voting refers to using the "winning" class as the final ground truth label. For example, if two raters say "yes" and one says "no," then the final label is yes (2/3). Similarly, if there are five raters, then the class obtaining three or more votes will be the final label, and so on. With an odd number of raters, the binary classification will always have a majority label when using majority voting.

Our research goal is to evaluate if majority voting is a justified strategy for binary classification when the task has a non-negligible degree of subjectivity (i.e., room for interpretation). To investigate this matter, we collect 3,000 ratings on 300 social media comments from a crowdsourcing platform and investigate how the dynamics of inter-rater agreement evolve when varying the number of raters.

We chose hate detection as the illustrative context for three reasons: (1) Prevalence of hate in online social media, (2) hate labeling has known issues of subjectivity and interpretation (Fortuna & Nunes, 2018; MacAvaney et al., 2019; Modha et al., 2020; Sood et al., 2012), and (3) there are several examples of studies (Davidson et al., 2017; Ibrohim & Budi, 2019; Magdy et al., 2015) applying the majority rule in crowdsourced labels to achieve ground-truth labels in this space.

Online hate detection is a growing field of research (see reviews in (Fortuna & Nunes, 2018; Waqas et al., 2019)) with broad cross-disciplinary interest among scholars from different communities, including HCI (Türkay et al., 2020). Typically, hate detection involves ML models with crowdsourced ratings as training data (Davidson et al., 2017; Mohan et al., 2017; Mondal et al., 2017; Salminen, Almerekhi, Milenković, et al., 2018; Waseem, 2016). A prominent

example is Perspective API (Alphabet, 2018), a tool by Jigsaw to score online comments for toxicity and hate. However, dataset quality is considered one of the most pressing challenges in online hate detection (MacAvaney et al., 2019; Modha et al., 2020; Vidgen & Derczynski, 2020). Our starting point for this study is that more research is needed into understanding the subjective nature of online hate and how this affects the training set creation process.

Related Literature

In HCI, crowdsourcing has been applied for various tasks, including taxonomy creation (Chilton et al., 2013), user studies (Kittur et al., 2008) such as graphical perception (Heer & Bostock, 2010), user interface performance (Komarov et al., 2013), accessibility (Hara et al., 2013), as well as generation of creative design outputs (Willett et al., 2012). Beyond HCI, social computing studies apply crowdsourcing to generate training samples for ML models (Davidson et al., 2017; Huang et al., 2014; Kocabey et al., 2018; Weber & Mejova, 2016).

ML techniques and crowdsourcing are a powerful combination for learning about online users. Nevertheless, the quality of the obtained annotations does not always lead to questions of dataset reliability (Alonso et al., 2013). Researchers tend to measure the inter-rater agreement as a proxy for quality (Alonso et al., 2015) to avoid such quality issues, with metrics such as Cohen's kappa (Cohen, 1960), Krippendorff's alpha (Krippendorff, 1980), Gwet's AC1 (Gwet, 2008), and others (Banerjee et al., 1999). When these metrics show a lack of agreement, several potential explanations arise. For example, guidance and instructions given to the raters may be inadequate or unclear (Pitkänen & Salminen, 2013), there may be fraudulent raters or bots (Peng et al., 2014) or there may be a sincere lack of attention (Alonso, 2015). A particular problem is *inherent subjectivity* (Salminen, Almerekhi, Dey, et al., 2018), meaning that the task actually has *no right or wrong answer*. To solve the issue of inherent subjectivity, researchers can deploy an odd number of raters and choose the last rater as a tiebreaker to assign the final label for the classified sample (Duwairi et al., 2014; Ibrohim & Budi, 2019; Magdy et al., 2015; Trieu et al., 2017; Volkova & Yarowsky, 2014).

Hate labeling is an example of a subjective labeling task. This is because individuals' opinions of what constitutes a hateful comment might differ despite the fact that a commonly accepted definition is provided (Alonso, 2015; Salminen et al., 2019; Salminen, Veronesi, et al., 2018). In (Salminen et al., 2019), the researchers analyzed 5,665 crowd ratings on 1,133 social media comments. The results indicated that individuals tend to agree on the extremes of a hate rating scale more than in the middle. The agreement was higher for comments that were, on average, considered less hateful and lower on comments that were generally rated as moderately hateful. The researchers suggest that this behavior helps reach an agreement on extreme cases (very hateful/not hateful at all) faster and more cost-

efficiently than obtaining an agreement on gray-area cases. In (Salminen, Veronesi, et al., 2018), the researchers collected 18,125 ratings from crowd workers in 50 countries, analyzing the effect of the country on the given hate scores. Even though geographic patterns were found, the conclusion is that hate ratings vary more by the individual raters than by countries.

There are many other subjective labeling tasks beyond hate detection. Examples include sentiment analysis (Cambria, 2016), peer-nominated personality ratings (Celli, 2011; Celli & Rossi, 2012), or virtually any topic dealing with opinions, attitudes, and preferences. The key distinction between subjective labeling tasks for ML applications of crowdsourcing is that they are conducted for the purpose of building a training set. This purpose tends to come with the explicit requirement of ground truth (Weber, 2015)—i.e., an assumption that the items have one true value. This is not the case for surveys, for which it is generally accepted (and expected) that the respondents' answers vary. In contrast, variation is a *problem* in a labeling task whose purpose is training set creation.

Method

Data Collection

We randomly sampled 300 comments from a previously published online hate dataset with known ground truth values (Salminen, Almerekhi, Milenković, et al., 2018). Half (n=150) of the comments are marked as hateful in the dataset, the other half as neutral. The crowd raters were recruited using the Appen platform (formerly known as CrowdFlower). The raters were presented with a simple binary decision task: "Is this comment hateful?" (Yes/No).

The raters were provided the following definition of hatefulness, similar to the definition applied by the dataset source (Salminen, Almerekhi, Milenković, et al., 2018): "*A hateful comment is rude, disrespectful or otherwise likely to make someone leave a discussion*."

We chose to have each sample labeled by ten raters. This choice is arbitrary, and we could also have aimed at having twenty or thirty raters as well. However, since this would have doubled or tripled the cost of acquiring data, we decided to choose ten raters. Overall, collecting 300 × 10 = 3,000 ratings was considered as enough data to examine the majority voting problem in crowdsourced labels. Also, according to the crowdsourcing platform's suggestions, eight test questions were created. For these test questions, we selected cases that were (according to our interpretation) clearly in either category (see Table 1). This decision is also in line with the best practices of the platform to make sure the test questions are as unambiguous as possible in order to avoid false rejection of workers, which will also result in extra cost for the crowdsourcing task.

Table 1. Test questions for the quality control of the crowdsourcing task. “Correct rate” indicates the share of correct crowd evaluations in the total. The observed correct rates are well aligned with the ground truth classes.

Comment	Class	Correct rate
Usain Bolt my hero has donated 10 million dollars to Haiti, all governments should hang their head in shame. A 100 metres runner did more than you. Bless up king Bolt	Non-hateful	92%
Chicago pizza is by far the best	Non-hateful	96%
Land ownership and the quiet title legal process in Hawaii are super complicated. For a better understanding of how it works, check out this resource from the Native Hawaiian Legal Corporation: http://www.nhlchi.org/resources/brochures/13#what-is-a-quiet-title-action	Non-hateful	95%
Fuck Obama and his Muslim apologism	Hateful	99%
he is not a fucking refugee he is a economic migrant fuck aj+	Hateful	100%
Stick your PROPAGANDA right up your FUCKING MUSLIM ASS!!!!!!!!!!!!!!!!!!!!	Hateful	98%
Fucking bullshit regressive channel. Aj+ fucking Muslim terrorist apologists. They won’t be singing in front of them when one of the Syrians or their decedents blow them up. I wonder what these Syrians think of Jews. I doubt they have a high opinion of them because their anti Semitic because their muslim	Hateful	99%
even stupid george bush knows its called quds force and not al quds, so fuck this channel and this whore.. this is a fucking jew channel, calling quds force al quds make it sound like al qaeda and al shabab and al nusra and those arab terrorist monkeys this is a persian force	Hateful	99%

Other quality control settings are also in line with the platform’s recommendations:

1. **Minimum Time per Page = 10 Seconds (Default).** This is the minimum time raters are required to complete a page of annotations. If less time is spent, the rater will be removed from the task.
2. **Disable Google Translate For Contributors = Enabled.** When enabled, this option disables Google Translate for raters using the Chrome browser to ensure that context and meaning are not lost in translation.
3. **Max Judgments per Contributor = Empty (Default).** This setting limits the maximum number of ratings that a rater can provide for the task. By default, the maximum ratings a rater can submit is limited by the number of test questions in the task. (In our case, eight test questions.)
4. **Quality Level** = 2 (“higher quality: a smaller group of more experienced contributors with a higher accuracy”).

The compensation for the workers was set at **rows per page = 5 (default)** and **price per page = USD 35 cents per page (default)**. These settings resulted in the **price per judgment = USD 7 cents (default)**. The parameters were set based on the belief that the platform’s defaults respect the minimum pay guidelines for crowdsourcing (Vaughan, 2017). The total cost for data collection was $316.26.

In other words, apart from the translation prevention and the increase of the quality level from default 1 to the higher level of 2, the other options were default.

We set the geographic targeting to the United States to gain some control over the cultural factors in the interpretation of hateful social media comments (Mubarak et al., 2017; Mubarak & Darwish, 2019).

Data Cleaning

A total of 300 social media comments were rated for hatefulness using crowdsourced ratings on a binary scale (yes/no). Each comment was rated by ten raters. The eight test questions that were rated by more than ten raters were excluded from the analysis for parsimony. Thus, the analysis comprised 292 comments with a total of 2,920 ratings. The ratings were sorted chronologically by creation date within each comment prior to the analysis.

Statistical Analysis

The statistical analysis was performed using the R software (v. 3.6.3). Counts and percentages were used to summarize the variables. The inter-rater reliability was assessed by using three measures: (1) percentage agreement, (2) Gwet's AC1 (Gwet, 2008), and (3) Krippendorff's alpha (K alpha). Using multiple agreement measures is advisable to ensure the consistency of the results (Cicchetti & Feinstein, 1990) by mitigating the impact of the shortcomings of any given metric on the overall findings. The AC1 and K alpha are chance-corrected agreement measures.

The K alpha measure can be used for nominal and ordinal outcomes and can take a value between 0 (perfect disagreement) and 1 (perfect agreement). It can also accommodate missing data, although in this case, we had none. The K alpha corrects the expected agreement by chance and can acquire lower values with high values of percentage agreement (Krippendorff, 1980).

The AC1 can be used when the expected agreement due to chance is high, which inversely affects the calculation of K alpha (Gwet, 2008). AC1 was developed as an alternative method in the presence of high expected agreement by chance, as it does not assume independence between raters. AC1 also supports categorical, ordinal, interval and ratio types of data and supports missing values.

Results

Number of ties based on the number of raters

We calculate the number of ties to understand how much using majority voting would affect the final labels. A tie is a situation where an equal number of raters think the comment is hateful and non-hateful (e.g., out of six raters, three choosing "yes" and three choosing "no" constitutes a tie).

Table 2 shows two important results. First, *there are no tie ratings, ever, when using an odd number of raters*. That is, even if there is an underlying tendency of disagreement among the raters, this can be obfuscated by choosing an odd number of raters and their majority decision on a given item.

Table 2. Ties for ratings with even raters

Raters	Number of ratings
2	79 (27%)
3	0%
4	37 (12.7%)
5	0%
6	26 (8.9%)
7	0%
8	16 (5.48%)
9	0%
10	12 (4.11%)

Second, *the proportion of comments with ties decreases with the increase in the number of raters*. Ties were observed for 79 (27%) and 37 (12.7%) comments when the ratings from the first two and four raters were used for the analysis, respectively. The number decreased to 26 (8.9%) when the ratings from six raters were used and further decreased to 16 (5.48%) when eight raters were used. The number of ties was lowest when all ten raters were used for the analysis (4.11%).

This finding can be interpreted, in a certain sense, as convergence to a consensus opinion on the "true" ratings of the items (see Figure 1). However, it is notable that *even with ten raters, there is a non-negligible number of ties*.

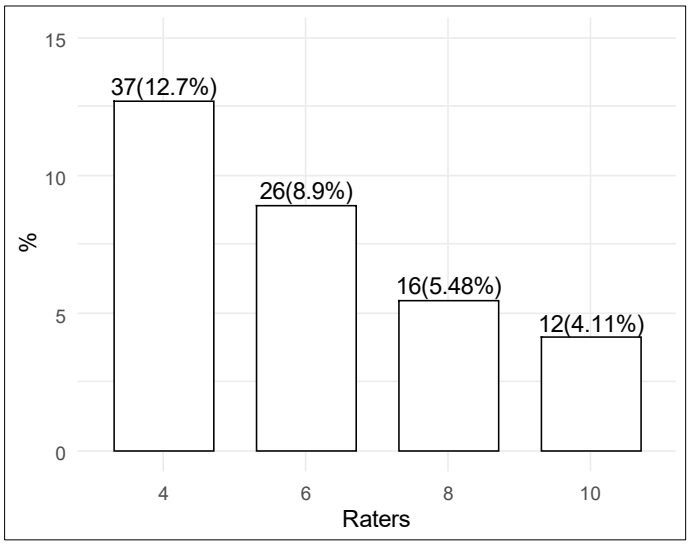


Figure 1: Ties based on the number of raters. The decreasing number indicates convergence to the "true" values. Yet, even with ten raters, some cases have equal support of being "yes" or "no."

Agreement on the Hatefulness of Comments

We next analyze the structure of the majority vote among the raters. This analysis is based on the majority rule, i.e., the comment was deemed hateful if more than 50% of the raters found it hateful. The proportion of comments that fit this criterion was calculated based on the number of raters. The analysis was performed for comments with an odd number of raters. Results in Table 3 show three interesting findings: First, the frequency of hateful comments corresponds well with the expected frequency (the ground truth had 150 comments labeled hateful, and the raters found 145-148 hateful comments). Second, the frequency of hateful comments remains stable from three to nine raters (at around 50%). Third, a full agreement among the raters (i.e., all of the raters agreeing that a comment is hateful) monotonically decreases from three raters (28.4%) to nine raters (19.5%).

Table 3. Structure of the majority vote on hateful comments.

Hateful comments (> 50%)		**Raters who found the comment hateful**							
	Raters	2	3	4	5	6	7	8	9
147 (50.3%)	3	64 (21.9%)	83 (28.4%)						
147 (50.3%)	5		30 (10.3%)	47 (16.1%)	70 (24%)				
148 (50.7%)	7			21 (7.19%)	26 (8.9%)	39 (13.4%)	62 (21.2%)		
145 (49.7%)	9				14 (4.79%)	17 (5.82%)	23 (7.88%)	34 (11.6%)	57 (19.5%)

According to the majority rule, with three raters, 100% of the comments have at least 66.7% agreement (2/3). However, what is the proportion of comments with five, seven, or nine raters with at least 66.7% agreement? Mathematically, the "worst" case for a given class to win decreases as the number of raters increases. For five, it is 3/5 = 60%; for seven, it is 4/7 = 57.1%; for nine, it is 5/9 = 55.6%. Our data shows that when the same comments are evaluated by five raters, 40.1% (n=117) of the comments have at least 66.7% agreement. With seven raters, 43.5% (n=127) have at least 66.7% agreement, and for nine raters, the value is 44.9% (n=131). These results imply that *using more raters helps understand the subjectivity of the task by teasing out differences in the agreement structure*.

Inter-rater Reliability

Our third analysis focuses on the agreement among the raters in all instances. The results in Table 4 show that the percentage agreement rate varies from 71.5% to 75.1%. The K alpha and AC1 measures were significantly different from zero, irrespective of the number of raters, as shown by the 95% confidence intervals

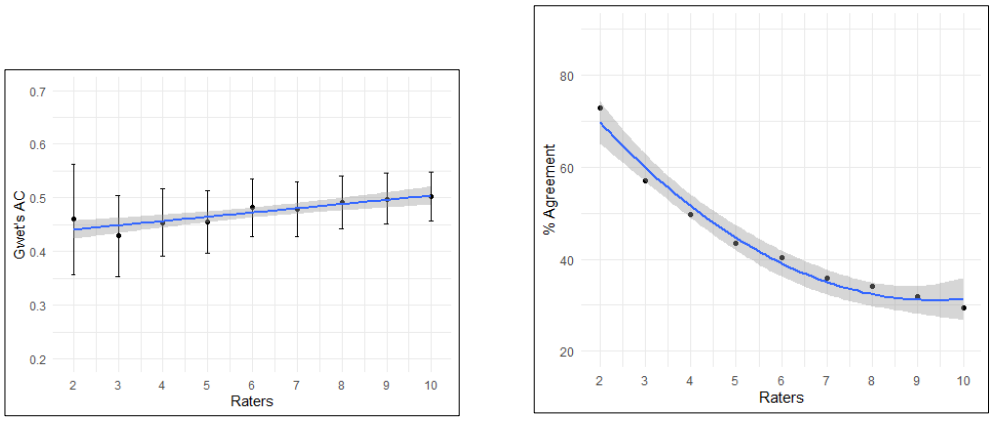
above 1. Interestingly, the agreement remains fairly stable throughout the increase in the number of raters. Ten raters have a higher agreement rate than any other number of raters, supporting the increase in the number of raters.

Table 4. Inter-rater reliability scores (95% confidence intervals in parentheses). The metrics show consistent results.

Raters	% agreement	K alpha	Gwet's AC
2	72.9%	0.459 (0.356, 0.561)	0.46 (0.357, 0.563)
3	71.5%	0.43 (0.354, 0.506)	0.429 (0.353, 0.505)
4	72.7%	0.455 (0.391, 0.518)	0.454 (0.391, 0.518)
5	72.7%	0.455 (0.397, 0.512)	0.455 (0.397, 0.513)
6	74.1%	0.481 (0.428, 0.534)	0.482 (0.428, 0.535)
7	73.9%	0.478 (0.427, 0.528)	0.479 (0.428, 0.53)
8	74.6%	0.491 (0.443, 0.539)	0.492 (0.443, 0.54)
9	74.9%	0.498 (0.451, 0.545)	0.498 (0.451, 0.546)
10	75.1%	0.502 (0.457, 0.548)	0.502 (0.457, 0.548)

The results in Figure 2a show that the 95% confidence intervals are overlapping, although the values tended to be slightly higher with the increase in the number of raters. Regression analysis was used to assess whether a statistically significant linear trend existed in the relation between the number of raters and AC1. Data points were weighted using the inverse of the standard error, so data points with a higher standard error (less confidence) had lower weight in the regression analysis. The results indicate a statistically significant positive linear trend ($B = 0.008$, $P < 0.001$). This indicates that increasing the number of raters is associated with a modest but significant increase in AC1.

Finally, a linear regression analysis shows a statistically significant quadratic trend (see Figure 2b) in the relation between the number of raters and the perfect agreement rate ($P < 0.001$) with a strong initial decline in the proportion of raters who were in perfect agreement and a slightly less strong association at later stages (after adding a 6th or 7th rater).



(a) **(b)**

Figure 2: **(a)** Gwet's AC1 based on the number of raters (the vertical lines represent the 95% confidence interval, and the horizontal line represents the regression line); **(b)** Perfect agreement based on the number of raters. The smoothed line represents the negative quadratic trend when increasing the number of raters.

Discussion and Practical Implications

In summary, using the majority vote tactic with three raters and binary classification is not recommended as the only option for building ML datasets, as this can cloud the subjectivity of the task and give a false sense of dataset validity. Even though the study only tested an online hate rating task, similar results are to be expected for other subjective rating tasks.

The results also imply that, while beneficial, increasing the number of raters cannot be considered the fundamental solution to the issue of agreement in subjective crowdsourcing tasks. Subjectivity can be so strongly ingrained in the data that no number of raters results in perfect agreement.

If uncertain, researchers can probe the subjectivity of their task by annotating a small sample of data with a large number of raters and observe how agreement and majority vote tendencies evolve. Another option is to sway from the requirement of one true label for every item in the ground truth. Instead, researchers can investigate the use of empirical distributions, as done in (Wulczyn et al., 2017). Essentially, whereas the one-true-label approach requires the predicted value to be either 1 (hateful) or 0 (non-hateful), the empirical description contains a tuple of values (e.g., [0.7, 0.3]). Hence, there is more information on the distribution of preferences. Depending on the number of classes and the readiness of the applied ML algorithm, empirical distributions can have a varying number of elements.

Previous research suggests that extremely hateful or non-hateful comments reach a consensus faster than comments in the mid-range (Salminen et al., 2019). Yet, we are not aware of any annotation schema that would leverage this property.

Some platforms such as Appen offer "dynamic judgments," a feature that collects *more* ratings for samples that struggle to reach consensus. However, what perhaps would be needed is the *exclusion* of such samples. If a sample is inherently subjective, collecting more ratings will not help resolve disagreements. Alternatively, these grey area comments could be labeled as such – e.g., apply label "indecisive" and use that as a third category for training the hate classifier.

While this analysis focused on hate detection datasets as the context, our findings apply to other training set creation tasks, e.g., those in the realm of NLP and sentiment analysis (Cambria, 2016; Celli, 2011; Celli & Rossi, 2012), as these fields generally face the same systematic issue of subjectivity.

Finally, we would like to point out that the are some general limitations when relying on crowd work for research purposes. For example, the lack of subject-matter expertise may be harmful to ML outcomes when the training data annotation would require specific domain knowledge (Alonso, 2015; Alonso et al., 2013). When recruiting crowd workers, this issue can partially be addressed by including training as a part of the annotation process (e.g., by using test questions that clarify where the crowd worker made a mistake), but this is not possible when the required level of expertise exceeds what can reasonably be trained in a short amount of time. Overall, researchers may benefit from expanding their views of how to design a crowdsourcing task, including questions about whether the ground truth unfolds as a result of a planned process, series of clarifications and redefinitions, or as a succession of surprises and repairs (Muller et al., 2021).

Conclusion

When tasks are subjective, using crowdsourced majority voting with three raters can hide real disagreements. Our results show that the rate of perfect agreement decreases with the increase in the number of raters. Researchers can label a small sample of their data with more than three raters (e.g., 5, 10) to validate the stability of their ground truth labels before conducting further analyses.

References

Almerekhi, H., Kwak, H., Salminen, J., & Jansen, B. J. (2020). Are These Comments Triggering? Predicting Triggers of Toxicity in Online Discussions. Proceedings of The Web Conference 2020, 3033–3040. https://doi.org/10.1145/3366423.3380074

Almerekhi, H., Kwak, H., Salminen, J., & Jansen, B. J. (2019). Detecting Toxicity Triggers in Online Discussions. *Proceedings of the 30th ACM Conference on Hypertext and Social Media (HT'19)*, 291–292. https://doi.org/10.1145/3342220.3344933

Alonso, O. (2011). Crowdsourcing for Information Retrieval Experimentation and Evaluation. In P. Forner, J. Gonzalo, J. Kekäläinen, M. Lalmas, & M. de Rijke (Eds.), *Multilingual and Multimodal Information Access Evaluation* (Vol. 6941, pp. 2–2). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-23708-9_2

Alonso, O. (2015). Practical Lessons for Gathering Quality Labels at Scale. *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1089–1092. https://doi.org/10.1145/2766462.2776778

Alonso, O., Marshall, C. C., & Najork, M. (2015). Debugging a Crowdsourced Task with Low Inter-Rater Agreement. *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, 101–110. https://doi.org/10.1145/2756406.2757741

Alonso, O., Marshall, C. C., & Najork, M. A. (2013, November 3). A Human-Centered Framework for Ensuring Reliability on Crowdsourced Labeling Tasks. *First AAAI Conference on Human Computation and Crowdsourcing*. First AAAI Conference on Human Computation and Crowdsourcing. https://www.aaai.org/ocs/index.php/HCOMP/HCOMP13/paper/view/7487

Alonso, O., & Mizzaro, S. (2012). Using crowdsourcing for TREC relevance assessment. *Information Processing & Management*, *48*(6), 1053–1066. https://doi.org/10.1016/j.ipm.2012.01.004

Alphabet. (2018). *Perspective API*. https://www.perspectiveapi.com/#/

Aroyo, L., Dixon, L., Thain, N., Redfield, O., & Rosen, R. (2019). Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions. *Companion Proceedings of The 2019 World Wide Web Conference*, 1100–1105. https://doi.org/10.1145/3308560.3317083

Banerjee, M., Capozzoli, M., McSweeney, L., & Sinha, D. (1999). Beyond kappa: A review of interrater agreement measures. *Canadian Journal of Statistics*, *27*(1), 3–23. https://doi.org/10.2307/3315487

Cambria, E. (2016). Affective computing and sentiment analysis. *IEEE Intelligent Systems*, *31*(2), 102–107.

Celli, F. (2011). Mining user personality in twitter. *Language, Interaction and Computation CLIC*.

Celli, F., & Rossi, L. (2012). The role of emotional stability in Twitter conversations. *Proceedings of the Workshop on Semantic Analysis in Social Media*, 10–17.

Chilton, L. B., Little, G., Edge, D., Weld, D. S., & Landay, J. A. (2013). Cascade: Crowdsourcing taxonomy creation. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1999–2008.

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology*, *43*(6), 551–558. https://doi.org/10.1016/0895-4356(90)90159-M

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Davidson, T., Bhattacharya, D., & Weber, I. (2019). Racial Bias in Hate Speech and Abusive Language Detection Datasets. *Proceedings of the Third Workshop on Abusive Language Online*, 25–35. https://doi.org/10.18653/v1/W19-3504

Davidson, T., Warmsley, D., Macy, M., & Weber, I. (2017). Automated Hate Speech Detection and the Problem of Offensive Language. *Proceedings of Eleventh International AAAI Conference on Web and Social Media*, 512–515.

Duwairi, R. M., Marji, R., Sha'ban, N., & Rushaidat, S. (2014). Sentiment analysis in arabic tweets. *2014 5th International Conference on Information and Communication Systems (ICICS)*, 1–6.

Fortuna, P. (2017). Automatic detection of hate speech in text: An overview of the topic and dataset annotation with hierarchical classes [Master's thesis]. Faculdade De Engenharia Da Universidade Do Porto.

Fortuna, P., & Nunes, S. (2018). A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, *51*(4), 1–30.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48. https://doi.org/10.1348/000711006X126600

Hara, K., Le, V., & Froehlich, J. (2013). Combining crowdsourcing and google street view to identify street-level accessibility problems. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 631–640.

Heer, J., & Bostock, M. (2010). Crowdsourcing graphical perception: Using mechanical turk to assess visualization design. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 203–212.

Huang, W., Weber, I., & Vieweg, S. (2014). Inferring nationalities of Twitter users and studying inter-national linking. *Proceedings of the 25th ACM Conference on Hypertext and Social Media - HT '14*, 237–242. https://doi.org/10.1145/2631775.2631825

Ibrohim, M. O., & Budi, I. (2019). Multi-label hate speech and abusive language detection in Indonesian twitter. *Proceedings of the Third Workshop on Abusive Language Online*, 46–57.

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. *Proceedings of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems*, 453–456. https://doi.org/10.1145/1357054.1357127

Kittur, A., Chi, E. H., & Suh, B. (2009). What's in Wikipedia?: Mapping Topics and Conflict Using Socially Annotated Category Structure. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1509–1512. https://doi.org/10.1145/1518701.1518930

Kittur, A., Lee, B., & Kraut, R. E. (2009). Coordination in collective intelligence: The role of team structure and task interdependence. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1495–1504.

Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., & Horton, J. (2013). The future of crowd work. *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*, 1301–1318. http://dl.acm.org/citation.cfm?id=2441923

Kocabey, E., Ofli, F., Marin, J., Torralba, A., & Weber, I. (2018). Using computer vision to study the effects of BMI on online popularity and weight-based homophily. *International Conference on Social Informatics*, 129–138.

Komarov, S., Reinecke, K., & Gajos, K. Z. (2013). Crowdsourcing performance evaluations of user interfaces. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 207–216. https://doi.org/10.1145/2470654.2470684

Krippendorff, K. (1980). Validity in Content Analysis. *Computerstrategien Für Die Kommunikationsanalyse*, 69–112.

MacAvaney, S., Yao, H.-R., Yang, E., Russell, K., Goharian, N., & Frieder, O. (2019). Hate speech detection: Challenges and solutions. *PLOS ONE*, *14*(8), e0221152. https://doi.org/10.1371/journal.pone.0221152

Magdy, W., Darwish, K., & Abokhodair, N. (2015). Quantifying public response towards Islam on Twitter after Paris attacks. *ArXiv Preprint ArXiv:1512.04570*.

Modha, S., Mandl, T., Majumder, P., & Patel, D. (2020). Tracking Hate in Social Media: Evaluation, Challenges and Approaches. *SN Computer Science*, *1*, 1–16.

Mohan, S., Guha, A., Harris, M., Popowich, F., Schuster, A., & Priebe, C. (2017). The Impact of Toxic Language on the Health of Reddit Communities. *Proceedings of the Canadian Conference on Artificial Intelligence*, 51–56. https://doi.org/10.1007/978-3-319-57351-9_6

Mondal, M., Silva, L. A., & Benevenuto, F. (2017). A Measurement Study of Hate Speech in Social Media. *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, 85–94.

Mubarak, H., & Darwish, K. (2019). Arabic offensive language classification on twitter. *International Conference on Social Informatics*, 269–276.

Mubarak, H., Darwish, K., & Magdy, W. (2017). Abusive language detection on Arabic social media. *Proceedings of the First Workshop on Abusive Language Online*, 52–56.

Muller, M., Wolf, C., Andres, J., Desmond, M., Joshi, N. N., Ashktorab, Z., Sharma, A., Brimijoin, K., Pan, Q., & Duesterwald, E. (2021). Designing Ground Truth and the Social Life of Labels. *Proceedings of ACM Human Factors in Computing Systems (CHI'21)*.

Peng, L., Xiao-yang, Y., Yang, L., & Ting-ting, Z. (2014). Crowdsourcing fraud detection algorithm based on Ebbinghaus forgetting curve. *International Journal of Security and Its Applications*, *8*(1), 283–290.

Pitkänen, L., & Salminen, J. (2013, November). Managing the Crowd: A Study on Videography Application. *In Proceedings of Applied Business and Entrepreneurship Association International (ABEAI)*.

Salminen, J., Almerekhi, H., Dey, P., & Jansen, B. J. (2018, October 15). Inter-rater agreement for social computing studies. *Proceedings of The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS - 2018)*. The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS - 2018), Valencia, Spain.

Salminen, J., Almerekhi, H., Kamel, A. M., Jung, S., & Jansen, B. J. (2019). Online Hate Ratings Vary by Extremes: A Statistical Analysis. *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, 213–217. https://doi.org/10.1145/3295750.3298954

Salminen, J., Almerekhi, H., Milenković, M., Jung, S., An, J., Kwak, H., & Jansen, B. J. (2018, June 25). Anatomy of Online Hate: Developing a Taxonomy and Machine Learning Models for Identifying and Classifying Hate in Online News Media. *Proceedings of The International AAAI Conference on Web and Social Media (ICWSM 2018)*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17885

Salminen, J., Veronesi, F., Almerekhi, H., Jung, S., & Jansen, B. J. (2018, October 15). Online Hate Interpretation Varies by Country, But More by Individual: A Statistical Analysis Using Crowdsourced Ratings. *Proceedings of The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS - 2018)*. The Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS - 2018), Valencia, Spain.

Sood, S., Antin, J., & Churchill, E. (2012). Profanity Use in Online Communities. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1481–1490.

Trieu, L. Q., Tran, H. Q., & Tran, M.-T. (2017). News classification from social media using twitter-based doc2vec model and automatic query expansion. *Proceedings of the Eighth International Symposium on Information and Communication Technology*, 460–467.

Türkay, S., Formosa, J., Adinolf, S., Cuthbert, R., & Altizer, R. (2020). See No Evil, Hear No Evil, Speak No Evil: How Collegiate Players Define, Experience and Cope with Toxicity. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13.

Vaughan, J. W. (2017). Making better use of the crowd: How crowdsourcing can advance machine learning research. *The Journal of Machine Learning Research*, *18*(1), 7026–7071.

Vidgen, B., & Derczynski, L. (2020). Directions in Abusive Language Training Data: Garbage In, Garbage Out. *ArXiv Preprint ArXiv:2004.01670*.

Volkova, S., & Yarowsky, D. (2014). Improving gender prediction of social media users via weighted annotator rationales. *NIPS 2014 Workshop on Personalization*.

Waqas, A., Salminen, J., Jung, S., Almerekhi, H., & Jansen, B. J. (2019). Mapping online hate: A scientometric analysis on research trends and hotspots in research on online hate. *PLOS ONE*, *14*(9), e0222194. https://doi.org/10.1371/journal.pone.0222194

Waseem, Z. (2016). Are you a racist or am i seeing things? Annotator influence on hate speech detection on twitter. *Proceedings of the First Workshop on NLP and Computational Social Science*, 138–142.

Weber, I. (2015). Demographic research with non-representative internet data. *International Journal of Manpower*, *36*(1), 13–25.

Weber, I., & Mejova, Y. (2016). Crowdsourcing health labels: Inferring body weight from profile pictures. *Proceedings of the 6th International Conference on Digital Health Conference*, 105–109.

Willett, W., Heer, J., & Agrawala, M. (2012). Strategies for crowdsourcing social data analysis. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 227–236. https://doi.org/10.1145/2207676.2207709

Wulczyn, E., Thain, N., & Dixon, L. (2017). Ex Machina: Personal Attacks Seen at Scale. *Proceedings of the 26th International Conference on World Wide Web*, 1391–1399. https://doi.org/10.1145/3038912.3052591

Yu, L., Kittur, A., & Kraut, R. E. (2016). Encouraging "Outside- The- Box" Thinking in Crowd Innovation Through Identifying Domains of Expertise. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, 1214–1222. https://doi.org/10.1145/2818048.2820025

Katerina Cerna and Claudia Müller (2021): Making online participatory design work: Understanding the digital ecologies of older adults. In: Proceedings of the 19th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies - Exploratory Papers, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591), DOI: 10.18420/ecscw2021_n22

Making online participatory design work: Understanding the digital ecologies of older adults

Katerina Cerna, Claudia Müller
University of Siegen
{katerina.cerna; claudia.mueller}@uni-siegen.de

Abstract. Participatory design (PD) is a meaningful approach to involve older adults into design; however, currently we lack understanding how to do such work online. In our paper, we report from a study where we organized 19 PD workshops online with older adults. We argue that to do so in a meaningful way, a mutually shaped understanding of older adults' digital ecologies is at the core of organizing such PD processes. We present an empirical account of how digital ecologies of our older participants have become an issue to tackle in the online PD workshops. Further, we provide a solution, a mapping technique, and report from our efforts to evaluate it, that should help to overcome the situation when digital ecologies become a problem in PD online.

Introduction

To make participatory design (PD) work, a certain type of work is necessary - as researchers we need to make the participation of everyone involved *work.* This is especially true when designing with and for communities that are not digitally well attuned and who need support in becoming able to participate in the PD process, such as for example older adults. Currently, PD is viewed as a meaningful way to involve older adults' into ideation and development of design concepts and digital technology. Traditionally, participatory design builds on the ideals of participation for everyone, involvement of heterogeneous stakeholders and the need for mutual learning (Kensing & Blomberg, 1998). In case of PD for and with older adults, mutual learning involves not only learning by the researchers about the stakeholders' situation to produce suitable technological solutions, but also about building the digital mastery of the older participants (Joshi & Bratteteig, 2016). CSCW and HCI literature provides a vast discussion on methods for including older adults in such processes, e.g interview methods, the usage of tool kits (Rogers et al., 2014), PD as a space for dialogue (Vines et al., 2012), cultural probes (Hensely-Schinkinger et al., 2018), and workshops (Pradhan et al., 2020). Locally employed measures for building empathy and mutual trust have been described as pivotal elements of PD workshops to make older adults feel comfortable in the overall situation, but especially also in their digital learning practices (Lindsay et al. 2012).

All these approaches build on the possibility to actually meet *on a physical site* and *in person*. However, meeting constantly in person might not be always possible. For example, Joshi and Bratteteig (Joshi & Bratteteig, 2016) show how participation was fluctuating due to the participants changing needs and capabilities. Also Vines et al.(2012) report that their older participants took part in only two to four particular workshops. It can be a matter of mobility, being sick or as we have seen recently due to COVID-19, but as a consequence it might not be possible to meet in person at all. But to build up the necessary digital skills for meaningful participation a regular attendance is necessary. Meeting online instead of on site provides a possible opportunity to tackle this problem. But to meet online, certain socio-material resources are necessary, such as personal devices (smartphones, laptops, stationary computers) with various applications or programmes. We understand this personal network as digital ecologies.

To our best knowledge, there is no PD involving older adults taking place online through video-conferencing only. This is not a surprise given that to meet in person to contribute to a design project is preferable for the participants as their main motivation is often social contact and learning in a group. Currently, we are hence lacking research on how to enable the older participants to participate in

online PD in a meaningful way. This study therefore focuses on the challenge of how to deploy a participatory design project aiming at co-creating didactic prototypes online with and for older adults. More specifically, we are interested in how older adults may be best supported in understanding their digital tools at home so that they have meaningful experiences from their participation in a series of 19 online workshops.

Through our collaboration with 20 participants we have learned that to organize online PD with older adults, a mutually shaped understanding of the respective digital ecologies is at the core of organizing such PD process. With our paper, we aim to contribute in the following ways. First, we would like to contribute with more facets from practice, how older people organize their personal digital ecologies and on sense-making processes in becoming participants in a collaborative and fully online-based project. With fleshing out particular instances from practice we would like to contribute to sharpening of the concept. Second, we propose a strategy how to map the digital ecologies of older adults for the purpose of online PD.

Related work

The ecological perspective on digital environments has been established in HCI for a long time (Blevis et al., 2015; Forlizzi, 2008). Using Gibson's ecological approach, Jung et al. (2008), for example, studied young students and how they manage personal digital artifacts interconnected to and woven into their lives. Ecologies of digital tools hold a potential to support collaboration and coordination of the users (Vasiliou et al., 2015). However, various studies also point out that engaging with ecologies of digital tools might be problematic as they are not static but dynamically evolve over time (Bødker & Klokmose, 2012). Another study emphasizes that in daily life communities where the ecologies tend to be more "messy" reflect the different preferences and competences of the communities (Bødker et al., 2017). The metaphor of an ecology of tools is useful in a number of respects. It draws attention to the way in which different devices are used for different purposes, to the fact that such an ecology implies a certain elegance and artfulness in use, and to the fact that too rapid changes in such ecologies create serious difficulties for certain populations.

To sum up, where the majority of studies of this kind have focused on the work context or on the so-called 'digital native' (Prensky, 2001), we, in contrast,

focus on older adults and how they deal with using their own digital devices, and specifically online collaboration tools when going online in situ. To our best knowledge, digital ecologies of older adults have not yet been explored, especially in the context of online PD.

Empirical setting and methods

In this section, we will first describe the method, followed by the empirical setting. This study reports from an ongoing participatory design project, whose goal is to enable older adults to become more autonomous in regards to digital tools. Methodologically, we want to reach this by co-creating a set of didactic digital prototypes together with the older adults which will promote and support learning of older adults and other relevant stakeholders. This project is part of a broader interdisciplinary effort called ACCESS focusing on fostering digital literacy of older adults.

Originally, the project was supposed to be taking place on-site, at the University and in local community facilities in a small German city. We recruited older participants from our existing research networks and through the local senior computer club: we started with 20 participants (11 women and 9 men). We invited the participants through our already existing social networks. Namely, we drew on our previously established collaborations with the local senior computer club. In addition, we also invited a group of older adults that was created for a previous research project. By drawing on already established connections, it was possible to continue developing our research work in a closer relation with the local communities. The participants were quite a heterogenous group, having various digital experiences and competences. In addition, three of the older participants who volunteer as instructors in the local senior computer club also took part in our PD project. The participants (including the instructors) were motivated in a different way to keep on engaging with our project: they wanted to keep on learning about new digital technologies, keep on collaborating with the university and engage with young people on a regular basis.

During the first on-site meetings, we installed a messenger application (Telegram) on their smartphones, which we used as the main coordination tool with the participants. Every participant joining our workshops already had at least one smartphone, although this was not a requirement. We were open and had some

smartphones to distribute to interested people as well as helping them to do the first steps with a digital device (Müller et al., 2015). Only one participant received a smartphone from us, however, she never took part in the online workshops.

As the global pandemic started to unravel during early 2020, we had to adjust and move with our workshops online. Even though it was not the original plan to conduct our empirical work online, it allowed us to explore how our participants made use of their digital devices and how they engaged online in making sense of their own practices in our joint collaboration sessions with tools which were familiar to them (smartphone or tablet) but also with new tools (such as Zoom or Miro) which we introduced.

Methodologically, we draw on an ethnographic-informed approach and on participatory design. We were interested in developing a rich understanding of the older adults' every-day practices so that we can build on them in our PD process (Randall et al., 2007). The work in the project is ongoing, but so far, we have conducted 24 workshops (5 onsite, 19 online). In addition, we have also conducted additional interviews as well as observations on-site when it was still possible (8 hours of observations in a local senior computer club). All online workshops were recorded, producing 50 hours of video materials. During the workshops, usually ten participants took part; accompanied by three moderators (German speaking) and one coordinator (not fully German speaking, first author).

We analyzed our data through an approach informed by Suchman and Trigg (1995). We rewatched the videos multiple times, searching first for problems the older adults experienced when participating in the online workshops; and we used our ethnographically-informed knowledge of the field during the analysis. One of the most visible problems was the issue of dealing with various devices to join the online session. We then searched for the particular instances of this problem in the whole corpus, which we then analyzed more in depth by drawing on elements of interaction analysis (Jordan & Henderson, 1995). By searching for themes in this chosen data, we then categorized these issues into themes: vocabulary, connection and possible activity. Even though we illustrate these issues by only one example, we could see a strong pattern of these issues in the data. All the transcripts were anonymized and all the names used in this text are pseudonyms.

Even though our research originally was not motivated by the global pandemic, it has had a big impact on our research activities. Majority of our participants were not familiar with video-conferencing tools, so we had to go through a learning phase for both us researchers (how to best support our participants?) and for our participants (how to use the video-conferencing tools?). Together, we tested out different tools (Jitsi, Skype and Zoom), and at the end

decided to use Zoom, because it provided us with the highest quality of video call in combination with necessary features, such as breakout rooms.

One of the key problems turned out to be not so much holding the sessions, but enabling older participants to reach the Zoom environment in the first place, as we had to instruct them how to use the technology *through* the particular technology. We have created and shared through Telegram manual-like resources with our participants (involving screenshots and written instructions). Despite that, this task turned out to be highly complex in some cases and we had to augment the initiation of the joint video calls with different tools of lower complexity, either texting through the common Telegram group or making individual phone calls.

Two key aspects of video-conferencing environments were heavily determining our online activities: Zoom is a *one-to-many communication channel.* Even though the whole group is present, it is only one person that can talk at once; often resulting into a pair of people having a conversation (instead of the whole group engaging), for example, to try solving a problem, and the rest of the group listening). Further, the *lack of access to the same physical space* was key (hence, for example, participants do not see what is happening on each others' screens). Both of these aspects change dramatically the way the researchers, instructors and peers can support each other (or not) in dealing with the digital tools and the online environment. These elements became more obvious, as we had the experience with the same group in working together in a shared physical site. When collaborating on site, it was possible to address the emerging issues with learning to work with digital tools in a one on one manner to address the highly individualized approaches of older participants. Sitting besides each other and pointing (with a finger) to specific parts of the phone or of the app were regular practices. The on-site workshops mostly had two parts: giving individual advice and help and talking about the workshop topic as the second part. Thus, the joint engagement between a person and a researcher on a device and about individual practices always was an important part of the on-site workshop with physical interaction and all senses involved (asking, listening, seeing, deictic practices (pointing "first click here, then here..."). This was in stark contrast to our online interaction. Especially one issue made the situations very complicated - which was that the researchers could not see with which devices and in which way the participants were trying to connect to the online sessions. Hence it was challenging to set up the online workshops on the basis of rather underexplored digital ecologies the older participants were acting in - as we will elaborate more below.

Understanding digital ecologies of older adults

In the following section, we will first describe the observed practices connected to digital ecologies of older adults, and illustrate the connected challenges in an engagement with a video tool of one of our participants displayed in an online workshop session. The empirical example illustrates that learning to use new tools can be challenging, especially when there is no support possible in presence. Second, we will present our own solution to this problem, a mapping technique, and the lessons learned from its evaluation.

The problem of digital ecologies

We encountered the issue of digital ecologies already during early stages of the online workshops, when we were testing out if we can use Zoom with our older participants. We have identified the following elements as constituting the digital ecologies of our older participants.

First, the older adults own a range of devices, such as smartphones, laptops and/or a tablet, with accompanying software. Each device involves their own "inner" ecology of interfaces, which the participants have to understand to be able to use the digital tools. Findings from our long-term online workshop observations indicate that the digital interfaces are all too similar for our older participants to distinguish them, and especially when switching between different programs (such as for example Zoom and browser) this becomes an issue as they cannot identify where exactly in their inner digital ecology they are. In addition, when joining the online workshops, we never knew in advance which of their devices the older adults will use and which will work during the particular session (sometimes for example sound would not work on their laptop and hence they would switch to a smartphone instead). This was again important information for us to have, as different devices and OS have also different interfaces and hence require different types of instructions.

Second element of digital ecologies are the individual practices of the participants. Even though it is conceptually problematic to distinguish between individual and group practices, as these are mutually shaping each other (Giddens, 1986), here we are using this term as a way to help us understand that older adults engage in a range of practices that differ from each other. For example, when it came to joining the Zoom session, which became the key step to be able to participate in the online workshops at all, the older adults have developed a range of different practices to do so. These practices involved for example typing the ID into Zoom, typing the link into the browsers, clicking the link in Telegram or in email. The different practices the older adults developed sometimes as a

consequence dealing with other contexts (for example, one woman struggled with joining our Zoom workshops, because we used a link instead of Zoom ID, as she was used to from her online choir). As we will see in the empirical example below, to have a range of practices like this is problematic for online PD, as it becomes difficult to support the older adults.

Third, another important part of the digital ecologies are the joint tools that we use to organize our PD online workshops. These involved Zoom (video-conferencing tool), Telegram (messenger) and Miro (white-board-like collaborative online tool). For the majority of our older participants these tools were new and they had to learn how to use them. The main difference from the individual devices and their software is that these tools would be the ones we were using to hold and coordinate the workshops and they were often an addition to the usual apps. Hence the older adults had to learn how to use these tools and we as a research team had to learn how to support the older adults in using them. The learning to use these tools often took the form of learning-by-doing.

To sum up, all these different combinations of devices, programs and (learning) practices form the digital ecologies of older adults. To make this all work together in the context of online PD, it is necessary to provide the older adults with support. However, to be able to provide them with support, we as researchers had to understand the particular ecologies. Since the older adults are often not aware of all the aspects of their own ecologies, it then becomes a challenge, how to incorporate it in the actual PD work. The following empirical example illustrates how the issue can develop during a particular session.

For example, some participants wanted to join the online workshops through their laptop; however, we first started to share the Zoom link through Telegram as that was our main communication channel with the participants. Majority of the participants had Telegram only on their smartphones and in turn did not know how to join the workshops with their laptops. This mismatch caused multiple troubles and we had to gradually establish new practices to attend this problem, for example, sending an email with the Zoom link, which the participants were able to access from their laptops. Talking about how one can join the Zoom session hence became a frequent topic at the beginning of the majority of the workshops, as it took several weeks before the joining practices became established. During a session in April 2020, when participants were prompted to reflect over the sessions, Monika takes a word:

Monika: So here's the thing, I tried with my cell phone to get this ID and the link in the first place and then I found out that with the PC and this ID it's much easier to get into the program than if I click on the email address and then mark it and then take it and whatever. That was quite simple, I wondered.

(both of the researchers' faces get a really confused look)
Monika: That's much easier than going into the email program and transferring the link and the ID and there... the link.
Researcher: Can you explain again how you did it in the end? I didn't quite understand that.
Monika: There was a number, I typed it in and then I was in.
Marvin: Oh, you typed the link that was in Telegram into the computer.
Monika: Exactly.
Marvin: Ah okay.
Monika: Yes, much easier than going into the email program and then transferring everything.

At this point researcher Marvin points out "You have Telegram, I think, yes, I think also on the computer", and as Monika confirms this statement, Marvin's continues: "Theoretically you can also do that, that if you open Telegram and then you should also just click on the link." Monika agrees with this utterance too.

We provide this example to illustrate how the mutual understanding of the digital ecologies was shaped through the moderators and older participants' interaction. In her account, Monika describes her preferred strategy on how to join the Zoom environment. As it is not very clear to the researchers what exactly she means, she explains further: instead of transferring the Zoom link from her email, she *types* the ID in a browser. As the workshop participants are trying to figure out how Monika actually enters Zoom, researcher Marvin points out that she also has Telegram on her laptop and hence can join the Zoom session from there, trying to point to a connection she missed.

The example illustrates several issues which are key to the problem of digital ecologies. First, there is the issue of ***vocabulary*** - to be able to make sense of the digital devices in the context of online PD, the participants and the researchers need to establish a common vocabulary. In the example above, we can see that Monika has not mastered the shared vocabulary yet, as she calls *link* an *ID* or says she clicked on an email address. Creating a common ground is a traditional aspect of PD (Kensing & Blomberg, 1998). However, in the context of online PD with older adults, mastering vocabulary is one of the first necessary steps to be able to create the common ground. Without the shared physical space, this step is key, as it is not possible to support the older adults in use by non-verbal gestures. In addition, as the possible range of terms is quite broad, the terms themselves are quite abstract (again, without the possibility to simply show the particular digital elements) and often Englishised hence make it more difficult for the older participants to remember.

Second, there is the issue of ***connection*** between the different devices and applications. In this example, Monika did not yet understand that since she had Telegram on both her laptop and her smartphone, she could access the Zoom meeting through Telegram message instead of using other ways. The only reason why the moderator Marvin knew that she had Telegram on her laptop is because they had installed it together several days before the workshop. Accessing one information through multiple devices is a common solution for example for digital nomads, who need to move seamlessly through their own digital ecologies and hence need a range of devices. However, in the context of online PD with older adults, this issue gains a new type of dimension, as this connection becomes a problem rather than a solution; as the researchers are at first not aware that Monika is missing the connection, they cannot provide her with appropriate support.

Finally, there is the issue of ***possible activity*** of the links (she does not have to copy a link in a way she finds complicated, she can click on it, which seems to be unclear to her); this is not only a problem of her not knowing *how to interact* with parts of the digital ecology; even though Monika says "ID", at that time, we actually did not provide our participants with the ID to enter the Zoom room. This example illustrates that the lack of Monika's understanding makes it difficult for the researchers to understand how she actually reached the Zoom room, and hence how to support it in case she would need help.

Were we successful in supporting Monika in her understanding? If joining the session does not become a problem, we actually do not know how the participants joined the session. However, during our workshop in August 2020, another participant was again struggling with accessing the Zoom room with her preferred device and was inquiring in the Zoom room about the necessary steps; on which Monika commented „Like I said, I just opened Zoom and clicked on the last Access Meeting and then I was already in". Here we can see that Monika has adopted the more common vocabulary as well as the established way to join Zoom. Even though she states that she clicked on a link in Zoom on the last meeting, that is not a common function in Zoom (and we assume that she meant link from Telegram). However, she did leave her previously established practice of typing information in and switched to a more common clicking on the link. To sum up, to be able to hold online PD workshops with our older participants, we needed to develop a mutually shaped understanding of the respective digital ecologies as they are at the core of organizing such PD processes.

Mapping technique development and evaluation

To address the problem of understanding the digital ecologies, we have developed a technique to map them. In this section we first describe the development of the technique and its consequential evaluation. As the mutually shaped understanding of digital ecologies builds on the interplay of learning how to use a particular tool by using it and the appropriate support learning-by-doing and self-directed learning was hence an important aspect to address when designing the mapping technique. In addition, we supported the older adults in their learning by providing necessary resources, such as visualizing elements of the digital ecologies and written and verbal instructions. To be able to mutually understand the digital ecologies of our older participants, we have developed a mapping technique that would support the process of mutually understanding the digital ecologies. By engaging in the collaborative process of mapping their own digital ecologies, we were aiming to support their own understanding of the connections among the devices, how to call particular parts of the ecologies and what activities are possible to do.

Practically, we chose to use a Miro board (a white-board like shared collaborative environment) because of its variability that is easy to adapt to our purpose; as well as the possibility to visualize the different connections. In addition, we already used Miro once two months before the evaluation workshop, which was a fun and an engaging session for our older participants. We structured the Miro space in a way so that participants could choose relevant elements relating to the particular challenges (an app/program, a device and an activity that could be done with an app, such as, joining a Zoom session) and then move it to their "spot" (Figure 1); gradually hence developing a map of their ecologies. We prepared both written and verbally delivered instructions for the older adults during the workshop. When introducing the task, we encouraged the older adults to navigate around in Miro by themselves, because during the last debriefing they expressed a wish to be allowed to do more tasks during our workshops autonomously; in addition this is also aligned with the overall aim of the project (supporting older adults in becoming more autonomous in regards to digital tools).

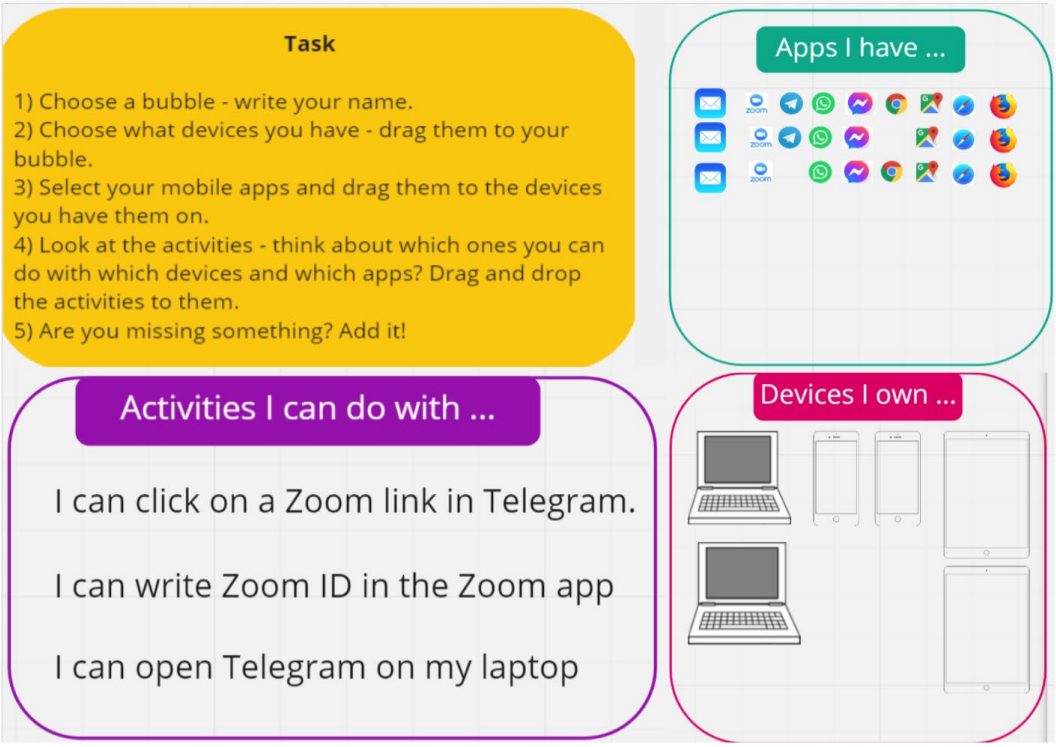


Figure 1: Section Task involves written instructions how to proceed; Section Apps I have involves icons of different applications; Section Devices I own involves different devices; Section Activities I can do with involves descriptions of different activities starting with "I can"

To understand how our prototype could be developed to be further appropriated by our older participants (and people outside of our PD context), we have dedicated one section of our online workshop to evaluate this particular aspect. The session was conducted in the same format as our regular online workshops that are through Zoom. However, this time Miro instead of enabling the understanding of the digital ecologies hindered their understanding of it.

We encountered troubles already at the beginning of the evaluation, when participants were "only" supposed to write their names into the Miro board. There were various reasons for this, for example that the current design of Miro does not change the cursor into the "typing" one, which is common in other contexts; hence making it difficult for the older adults to know where they are supposed to start typing. This seemingly trivial problem took more than 20 minutes of the session moderator trying to solve it; which often involved trying to attend to three participants having a problem at the same time. Overall, this also set up the task in a very confusing way.

Another challenge was to be able to recognize which application on their laptops the participants actually see. As we describe above, this is a common problem connected to the inner digital ecology of older adults. Leaving Zoom and switching to another program such as a browser is often a problematic moment, as the participants can "get lost" in their own ecology. As we can only rely on the older adults' verbal description of where they are, it becomes a challenge especially when dealing with new tools such as Miro.

After 20 to 30 minutes of trying to support the older adults in their own activities, we decided to change the strategy to be able to proceed with the session. Instead of the older adults exploring Miro and mapping their ecologies themselves, the session moderator shared his screen and started to build the ecologies based on the older participants' instructions. Through this approach, we were able to map the digital ecologies of the older adults in a way that maybe was not so "self-exploratory" as we wanted it to be but on the other hand it became possible to engage in the activity for everyone, not just the digitally more advanced. The final product can be seen in Figure 2.

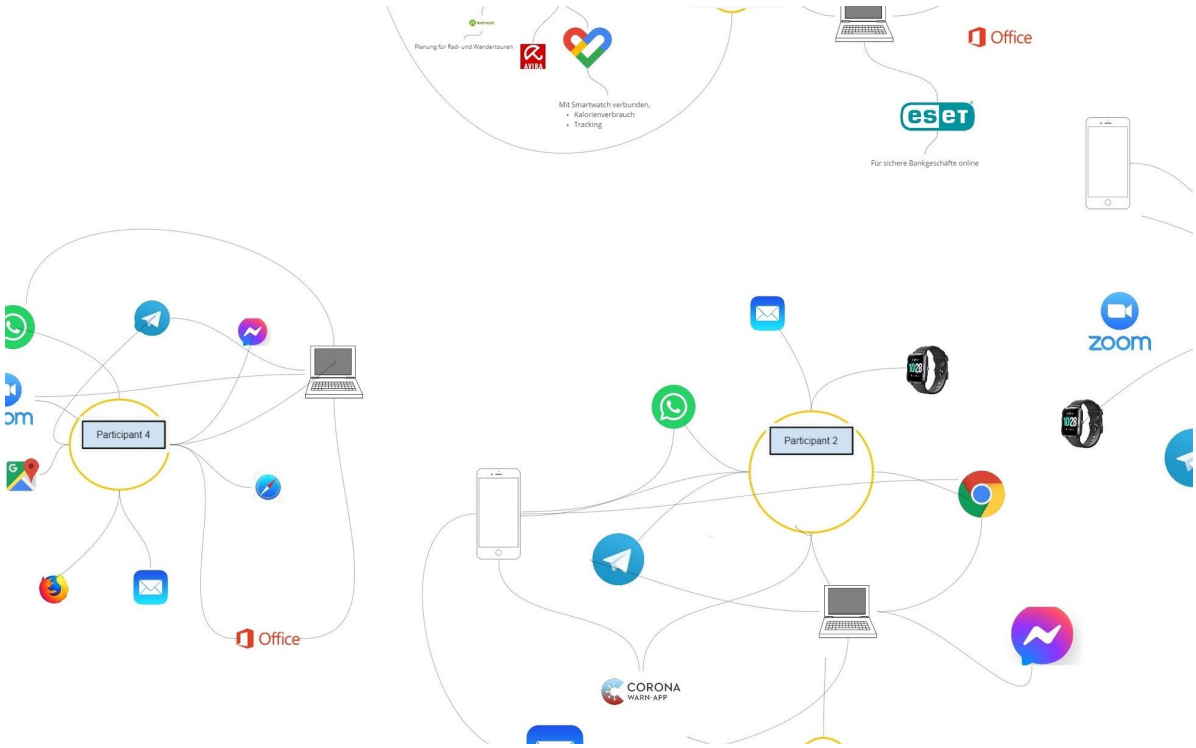


Figure 2: Final overview of the digital ecologies mapping

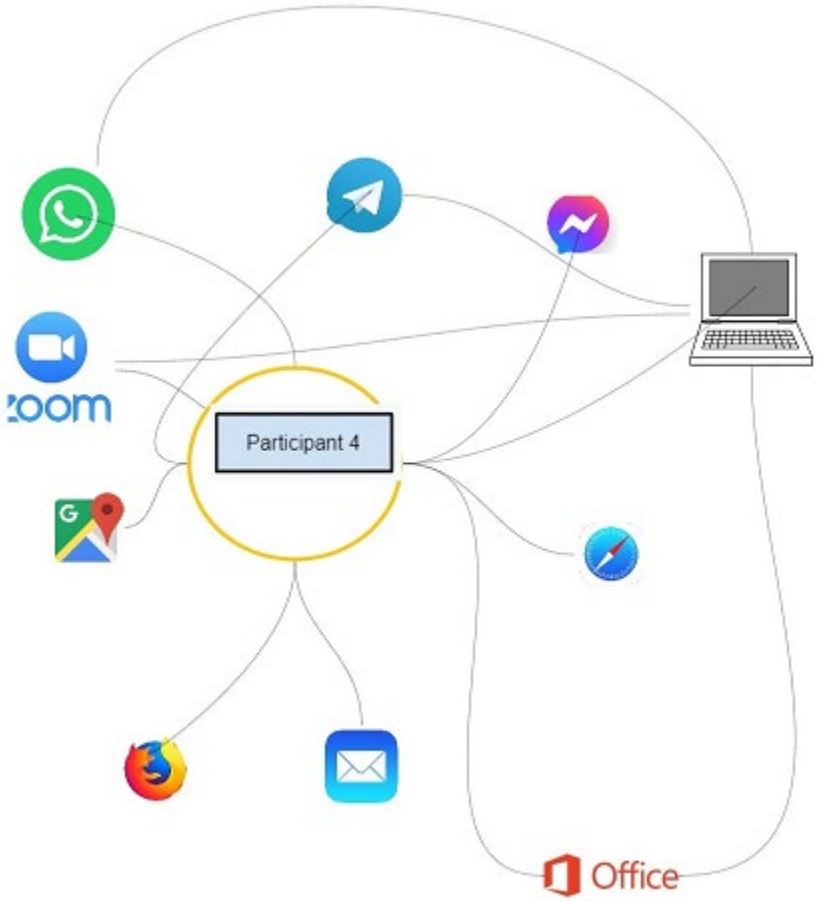


Figure 3: Zoomed-in excerpt from the digital ecologies mapping

After the session, during the debrief, participants expressed unclarity and confusion about the task, especially related to the speed of the instructions "Miro is too intense", which is an evaluation which we did not hear during the previous Miro mediated session. On the other hand, the process also contributed to some extent in what we were hoping for; one of the participants for example said: "Miro was my biggest issue, but I was impressed by the mapping of the ecologies, about how much apps I and others use".

To sum up this section, we have learned that to support mutually shaped understanding of the particular digital ecologies of older adults, it is possible to map them, however if the following is included.

1) If exploring the digital ecologies through another digital tool, the researchers need to consider that the possible change to the existing digital ecology might be a challenge. This enrichment should be gradual and in pace suitable to the older participants.
2) It is necessary to consider an adequate support for the older adults to navigate in their inner ecologies, when switching between the different tools. This might involve teaching older adults how to share their screen or taking photos of it and share them in a common communication channel.
3) It is key to reflect together on the final result. This step for us got lost because of the early difficulties, however even debriefing together at the end showed us some things our participants have learned (or not) from the mapping.

Concluding remarks

To sum up, in our paper we have presented how the digital ecologies of older adults can pose a challenge when involving the older adults into online PD. Variety of devices and programs, which are connected in various ways; individually established practices from a variety of contexts and jointly used tools which are introduced and used through error-trial and learning by doing for the older adults, they all form together the digital ecologies of the older adults. The digital ecologies of older adults are not a problem *per se*. However, since the online workshops build on involvement of new tools into the existing digital ecology of them, they will need support when accessing the online workshops. To be able to provide them with meaningful support, we as researchers need to understand which devices the older adults use, which program they see on their screens and which individual practices each participant is used to. This creates a complex environment, which is not automatically obvious at a first glance. More importantly, the only way to understand it is through supporting the participant in their own understanding of the digital ecologies. Without this understanding they will not know which parts of the ecologies are relevant for the online PD context and in turn possibly how to use them. In other words, understanding of digital ecologies is not a one-way process, but emerges from activities that mutually shape each other.

A question also is if the above described challenges are common for older adults or connected to the novelty of described technology. We believe that the possible explanation is somewhere in between: the older adults did not struggle with the digital tools because of their age, but rather because of their particular

needs. These needs are formed by the older adults' life and learning trajectories - many older adults did not use digital tools at work, as well as many of them are not used to the learning-by-doing approach. As a consequence, it was for example difficult for them to distinguish the various apps on their devices. Further, the designers of the involved digital tools do not consider some changes which are common for aging, such as troubles with sight (many of our participants struggled with the small size of the buttons) or less sensitivity in fingers or the heterogeneous interests and practices of this group.

Through our empirical work, we have learned that to organize online PD with older adults, a mutually shaped understanding of the respective digital ecologies is at the core of organizing such PD processes. More specifically, the researchers and participants can make the online PD work through the following activities:

- Developing a common vocabulary as the first step that helps everyone involved to orient themselves in digital ecologies.
- Understanding how the different devices are connected through installed applications and programs.
- Support the participants in learning the different activities they can do to join and participate in the online PD.

In addition, we have also explored ways how to overcome the challenge of digital ecologies in online PD with older adults. When mapping the digital ecologies of older adults, it is key to consider:

- which additional technologies one is adding to the existing digital ecologies by the mapping,
- adequate support for "inner" ecology navigation,
- and reflect over the mapping process and the final result of the mapping.

To conclude, to make online PD with older adults work, a specific type of work is necessary. This type of work needs to be focused on mutually shaped understanding the digital ecologies that the older participants use to join the online activities. In our paper, we propose a mapping technique that can be used to deepen the understanding of digital ecologies and hence secure a meaningful participation in the design process for the older adults.

References

Blevis, E., Bødker, S., Flach, J., Forlizzi, J., Jung, H., Kaptelinin, V., Nardi, B., &

Rizzo, A. (2015). Ecological Perspectives in HCI: Promise, Problems, and Potential. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 2401–2404.

Bødker, S., & Klokmose, C. N. (2012). Dynamics in artifact ecologies. *Proceedings of the 7th Nordic Conference on Human-Computer Interaction Making Sense Through Design - NordiCHI '12*. the 7th Nordic Conference, Copenhagen, Denmark. https://doi.org/10.1145/2399016.2399085

Bødker, S., Lyle, P., & Saad-Sulonen, J. (2017). Untangling the Mess of Technological Artifacts: Investigating Community Artifact Ecologies. *Proceedings of the 8th International Conference on Communities and Technologies*, 246–255.

Forlizzi, J. (2008). The Product Ecology: Understanding Social Product Use and Supporting Design Culture. *International Journal of Design*, *2*(1), 11–20.

Giddens, A. (1986). *The Constitution of Society: Outline of the Theory of Structuration* (p. 402). University of California Press.

Hensely-Schinkinger, S., Schorch, M., & Tellioglu, H. (2018). Using Cultural Probes in the Sensitive Research Setting of Informal Caregiving. A Case Study. *I-Com*, *17*(2)), 103–117.

Jordan, B., & Henderson, A. (1995). Interaction Analysis: Foundations and Practice. *The Journal of the Learning Sciences*, *4*(1), 39–103.

Joshi, S. G., & Bratteteig, T. (2016). *Designing for Prolonged Mastery. On involving old people in Participatory Design*. https://www.semanticscholar.org/paper/d5dd476c28cc8e2e06ed870d3d400c

9e94bc23b0

Jung, H., Stolterman, E., Ryan, W., Thompson, T., & Siegel, M. (2008). Toward a framework for ecologies of artifacts: how are digital artifacts interconnected within a personal life? *Proceedings of the 5th Nordic Conference on Human-Computer Interaction Building Bridges - NordiCHI '08*. the 5th Nordic conference, Lund, Sweden. https://doi.org/10.1145/1463160.1463182

Kensing, F., & Blomberg, J. (1998). Participatory Design: Issues and Concerns. *Computer Supported Cooperative Work: CSCW: An International Journal*, *7*(3), 167–185.

Müller, C., Hornung, D., Hamm, T., & Wulf, V. (2015). Practice-based Design of a Neighborhood Portal: Focusing on Elderly Tenants in a City Quarter Living Lab. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2295–2304.

Pradhan, A., Jelen, B., Siek, K. A., Chan, J., & Lazar, A. (2020). Understanding Older Adults' Participation in Design Workshops. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–15.

Prensky, M. (2001). Digital natives, digital immigrants. From On the Horizon. *MCB University Press*, *9*(5), 1–6.

Randall, D., Harper, R., & Rouncefield, M. (2007). *Fieldwork for Design: Theory and Practice*. Springer, London.

Rogers, Y., Paay, J., Brereton, M., Vaisutis, K. L., Marsden, G., & Vetere, F. (2014). Never too old: engaging retired people inventing the future with MaKey MaKey. *Proceedings of the SIGCHI Conference on Human Factors*

in Computing Systems, 3913–3922.

Suchman, L. A., & Trigg, R. H. (1995). Understanding Practice: Video as a Medium for Reflection and Design (Excerpt). In R. M. Baecker, J. Grudin, W. A. S. Buxton, & S. Greenberg (Eds.), *Readings in Human–Computer Interaction* (pp. 233–240). Morgan Kaufmann.

Vasiliou, C., Ioannou, A., & Zaphiris, P. (2015). An Artifact Ecology in a Nutshell: A Distributed Cognition Perspective for Collaboration and Coordination. *Human-Computer Interaction – INTERACT 2015*, 55–72.

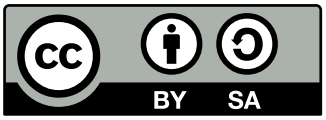
Vines, J., Blythe, M., Dunphy, P., Vlachokyriakos, V., Teece, I., Monk, A., & Olivier, P. (2012). Cheque mates: participatory design of digital payments with eighty somethings. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1189–1198.

Midas Nouwens & Clemens Nylandsted Klokmose (2021): A Survey of Digital Working Conditions of Danish Knowledge Workers. In: Proceedings of the 19th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies - Notes, Reports of the European Society for Socially Embedded Technologies (ISSN ISSN 2510-2591), DOI: 10.18420/ecscw2021_n24

A Survey of Digital Working Conditions of Danish Knowledge Workers

Midas Nouwens & Clemens Nylandsted Klokmose
Digital Design & Information Studies, Aarhus University
midasnouwens@cc.au.dk, clemens@cc.au.dk

Abstract. We present a representative survey of the digital working conditions of 466 Danish knowledge workers. We provide data on 1) the hardware and software they use to accomplish their main job activities, 2) the strategies they use to personalise their software, and 3) their digital competences. Our results show that the average Danish knowledge worker primarily uses a laptop and a smartphone to accomplish their work; they use an average of four software applications, mostly developed by large US corporations; they infrequently personalise their software using built-in settings and rarely personalise using plugins, scripts, or reprogramming; they are most capable in using collaboration and communication tools, feel more comfortable formatting other worker's digital content than creating their own, and are confident they can solve most technical issues. These results put into question the relevance of the long-standing Personal Computing dream envisioned by HCI pioneers, highlights the tensions between software applications and the digital sovereignty of the European continent, and emphasise the importance of including digital tools in our conceptualisation and regulation of working conditions.



Introduction

Knowledge workers are the prototypical professional software users, and continue to be the occupational category whose work activities are most supported (or at least mediated) by computers (Bughin et al., 2016). Knowledge workers have also been one of the main user groups studied by HCI researchers, and CSCW in particular. The practice-oriented research program of CSCW, however, has emphasised understanding particular contexts of computer use through interview and observation methods, and limited use has been made of probability-based social surveys (Wallace et al., 2017). Consequently, despite being a centrally recurring figure in HCI studies, little is know about the structural characteristics and conditions of knowledge workers who use computers to accomplish their daily work activities. Such an understanding can help establish generalisable knowledge about computer supported knowledge work and allows us to make informed prioritisations about which issues and communities to focus on using the more traditional interview and observation methods of CSCW research.

This study contributes a representative survey of digital working conditions of Danish knowledge workers – the most digitalised industry in one of Europe's most digital countries (European Comission, 2020). Thematically, this topic was operationalised through the following three sub-questions:

- What hardware and software do knowledge workers use to accomplish their main job activities?
- What strategies do knowledge workers use to personalise their software?
- What level of digital competences do knowledge workers have?

Using the answers to these questions, we paint a portrait of the digital characteristics and working conditions of knowledge workers in Denmark, which can help inform small-sample studies on the impact of digitalisation and discussions about the direction of digital policy to mitigate digital harms.

Background

Informational Capitalism

In the last half century, the political economies of most OECD countries have been transforming from *industrial* capitalism to *informational* capitalism. This qualifying adjective to capitalism follows Castell's seminal "The Rise of the Network Society" (2009), in which he augments the Marxist concept of a society's *mode of production* (capitalism, feudalism) with the idea of a *mode of development* (industrialism, informationalism). The mode of development tries to explain how the same mode of production can have different levels of surplus by identifying what the fundamental element is that increases productivity. In industrial capitalism, Castell argues, the main productive elements are new sources of energy (e.g., steam, electricity, oil) and how effectively they are used throughout production and distribution processes. In informational capitalism, the main source

of increased productivity comes from the use of "the technology of knowledge generation, information processing, and symbol communication" (Castells, 2009, p. 17)[1]. As the economy shifts its orientation from energy to information as the primary source of surplus value, the creation, accumulation, and use of that information become the organising principles for capitalist activity.

The Knowledge Economy

The informational capitalist system underpins the emergence of the knowledge economy in the 1990s-2000s: an economic structure whose largest share of growth comes from using knowledge to produce goods and services. Motivated by the slowing down of capital returns on mass-produced physical goods and increase of global competition, many countries committed to the idea of "knowledge" as the new, more efficient asset that would guarantee continued economic growth (examples of knowledge-based capital include patents, intellectual property, brand-equity, innovation research, and, of course, software). We can observe this shift concretely through the policy agendas of the European Commission. In 2000, formalised in the "Lisbon Strategy", the European Union committed itself to the idea of the information society and aimed to make the EU "the most competitive and dynamic knowledge-based economy" (European Parliament, 2000). In the following "Europe 2020" agenda set out ten years later, it repackaged that aim as the "digital economy", with initiatives such as the Digital Agenda, the Digital Single Market, and the Grand Coalition for Digital Jobs and Skills. The goal was to create an economy which could "exploit the potential of Information and Communication Technologies in order to foster innovation, economic growth and progress" (European Commission, 2010).

The Knowledge Worker

In knowledge economies, the *knowledge worker* – provocatively called "human capital" – has become the most in-demand commodity, as a large share of the surplus value is assumed to be created when the worker has more knowledge and uses it more effectively. It should be noted that the concept of knowledge work suffers from policy evangelism and lacks an operationalised definition. EU and OECD white papers have variously attempted to capture knowledge work by describing it based on the sector or industry they work in, the activities common in their work, the level of education required, or their occupation category, but none have allowed governments and businesses to measure and intervene effectively in this type of labour (See Brinkley et al., 2009 for a discussion). At the most abstract level knowledge work refers to any work that uses existing information in flexible and innovative ways to produce new information from which value can be

[1] Castell acknowledges that information plays an important role in other modes of development (and production) as well, but argues that the key difference in informationalism is that surplus is created through the application of information on information itself: knowledge is used to increase the quality and production of knowledge, rather than, say, the production of material goods.

extracted. One of the core goals of the European Commission's policies, then, has been to increase the share of knowledge workers in the European labour force. Initiatives have focused on raising the average level of education of the labour force, increasing the share of women in the labour force, and creating opportunities for workers to re-/upskill their digital competences.

The mediating role of digital technologies

The story of the knowledge economy, the "knowledge worker" as an occupation, and digital technologies are deeply connected, stretching back half a century. The application as a model of software first emerged during the late 1970s and early 1980s in the United States, and in large part became a commercially successful mass-market product because it managed to capture the imagination of large corporations and white-collar office workers (Nouwens, 2020). One reason why the knowledge economy became a viable alternative to the manufacturing economy was because computers increased the rate at which information could be produced and processed by orders of magnitude, and because increasingly user-friendly application software made it possible for workers to leverage that capability at scale. The knowledge worker as an occupation continues to be tightly coupled with the effective use of applications as the main tools of production (Nouwens and Klokmose, 2018). To this day, the more knowledge intensive industries continue to be the most digitised (Bughin et al., 2016).

The connection between knowledge work and software design is also foundational to the field of Human-Computer Interaction; the much-venerated line-up of North-American computing pioneers all imagined computers as empowering knowledge tools. Bush (1945) described his Memex as a device that would be an "enlarged intimate supplement to [a person's] memory". Licklider (1960) dreamed of a man-computer symbiosis where "the resulting partnership will think as no human brain has ever thought". Engelbart (1962) believed that "man's[2] problem-solving capability represents possibly the most important resource possessed by a society", so his Augmenting Human Intellect projects tried to increase "the capability of a man to approach a complex problem situation, to gain comprehension to suit his particular needs, and to derive solutions to problems". Kay (1990) called computers intellectual amplifiers that "would actually change the thought patterns of an entire civilization". Kay and Goldberg (1977) imagined software as a "clay of computing" that would let future knowledge workers use the computer to "mold and channel its power to his own needs".

Digital technologies continue to play a central role in the sociotechnical imaginaries surrounding the "Future of Work". If this future is digitally mediated, any analysis of labour-related concerns (e.g., employer-employee relationships, job quality, de/re/upskilling) will now have to consider what role the software's design, development, and deployment plays in the labour process. Here, HCI research and

[2] And, we assume, also other genders.

non-academic studies have left a gap. White papers and policy initiatives by governmental, non-governmental, and commercial research institutes have focused almost entirely on data and skills as the two main components for a digital, globally competitive economy, but curiously ignore the computational tools that workers use on a day-to-day basis to productively leverage those data and skills. HCI research at large takes software design as one of its core subject matters, but its methodological focus on practice-based studies (Wallace et al., 2017) leaves us in the dark about the larger, structural conditions of digital work. This study aims to provide such an initial understanding by contributing a (relatively) large-scale, representative survey of the digital working conditions of knowledge workers in Denmark.

Method

Instrument design

The survey consisted of a mix of 18 open and closed questions, with a possible maximum of 24 questions depending on specific conditional answers. The first question of the survey was used to filter respondents based on their occupation, using the sub-major groups of the 2008 version of the Danish International Standard Classification of Occupations (DISCO-08). The rest of the survey was divided into two sections: one with questions about the respondents' use of digital technologies, and one about their demographic characteristics.

The section about digital technologies consisted of questions about the hardware and software they used to accomplish their work activities (which and how many devices, what operating systems, and which software applications for each device); about whether they adapted their software (how often, and using which strategy); and about their digital competences (e.g., digital communication, collaboration, problem solving). The question regarding software adaptation was conceptually informed by partially-overlapping taxonomies developed by Mørch (1997), Trigg et al. (1987), and MacLean et al. (1990), resulting in four adaptation strategies: using the software's built-in preference settings, through plugins or add-ons, using scripts or macros, and by reprogramming the source code. The questions regarding digital competences were based on the self-assessment survey of the European Commission's Digital Competence Framework, where a participant can rank their competence level (basic–intermediate–advanced) for different skill categories (e.g., information processing, content creation, problem-solving) (see Carretero et al., 2017 for the full scale).

The section about demographic characteristics included questions about employment status (e.g., full-time, self-employed, unemployed, retired), job title, primary work activities, sector (public, private), and industry (e.g., financial and insurance, education, construction). The industry categories were based on the second revision of the Statistical Classification of Economic Activities in the European Community (NACE rev 2) (EUROSTAT, 2008). NACE is a multi-level

classification with 21 first level categories, each of which is further broken down into more specific activities. This study used 14 of the top level categories, and a selection of the second level classifications of 5 other categories. Two categories (sections T and U) were not included.

Data collection

Procedure

The data was collected between July 12 and 22, 2018 by YouGov, a global internet survey and data analytics company which maintains a panel of respondents across multiple demographic characteristics. Respondents earn points for completing surveys which can be exchanged for cash, vouchers, or prize draws.

Participants

A total of 3944 respondents between the ages of 18 and 74 were contacted, with quotas on gender, age, and region to reach a nationally representative sample.

		Sample		**Population**
		3945	*100%*	*100%*
Gender	*Female*	2148	54,4	49,8
	Male	1797	45,6	50,2
Age	*18-34*	866	22	30
	35-54	1637	31,5	37,4
	55-74	1442	36,6	32,6
Region	*Capital city*	1260	31,9	31,9
	Sjælland	577	14,6	14,4
	Syddanmark	813	20,6	20,9
	Midtjylland	863	21,9	22,6
	Nordjylland	432	11	10,2

Table I: Unweighted, unfiltered sample and overall population distribution

Variables

In addition to the data gathered through the survey instrument described in *Instrument design* on p.5, the YouGov service included pre-existing background information about the respondents gender, region, age, civil status, and education. The gender data was binary (male, female), and level of education followed the 2015 version of the Danish International Standard Classification of Education (DISCED-15).

Data processing

The sample was cleaned to increase the data quality and processed to make it representative.

The data was cleaned based on 1) occupation, 2) non-response, 3) qualitative data quality, and 4) overall response time. 2474 respondents were screened out because their self-reported occupation did not match our definition of knowledge work (i.e., not falling in the DISCO-08 categories of managers, professionals, and technicians and associate professionals[3]). 450 respondents were removed because they did not complete the survey. Respondents were asked to report which software applications they used to accomplish their main job activities per type of device (laptop, desktop, tablet, smartphone). This qualitative data was processed using fuzzy matching algorithms in OpenRefine[4] and manual inspection, resulting in a standardised list of software. All unreasonable answers (e.g., "asdfghjkl", "none") and software names that could not be identified were replaced with the value "-1". All participants with this response for any single, device-related software question were removed from the data set (n = 525). The median response time for the survey was 7 minutes, with the first quartile at 5 minutes and the third at 10 minutes. All respondents with a response time below 2,5 minutes and above 30 minutes were removed (n = 29). After the cleaning, the final sample size corresponds to 466 knowledge workers.

Post-stratification weights were applied to correct for non-responses using the marginal distribution of occupation category separated into sex (female, male) and sector (public, private). Information about the population was retrieved from Danmark Statistik, the official statistics bureau of the Danish government, specifically from "LONS20: Earnings by occupation, sector, salary, salary earners, components and sex"[5]. The weights were calculated using Iterative Proportional Fitting (IPF). Briefly, IPF is a method that forces the marginal distribution of a sample to match those of the population by applying a weight to each individual row. It does this by fitting the sample to the population using one demographic statistic at a time (e.g., gender). Once completed, it does the same for the next statistic, until the final distribution equals the population's.

The answers regarding device operating system had to be removed because of a flaw in the conditional logic of the survey that meant respondents were inconsistently shown the question.

---

[3] "Ledelsesarbejde" i.e., "Management"; "Arbejde, der forudsætter viden på højeste niveau inden for pågældende område" i.e., "Work which requires the highest level of knowledge for the field concerned"; and "Arbejde, der forudsætter viden på mellemniveau" or "Work that requires intermediate level knowledge".

[4] http://openrefine.org/

[5] Available here: https://www.dst.dk/en/Statistik/dokumentation/documentationofstatistics/structure-of-earnings

Results

The results are divided into the sections Hardware Working Environment, Software Working Environment, Digital Competences, and Digital Appropriation Strategies.

Hardware Working Environment

Contemporary knowledge workers have a variety of digital devices to choose from to perform their tasks, ranging from more traditional desktop computers, now-common laptops and smartphones, to the still fledgling tablet form factor.

The survey results indicate that the laptop and smartphone are by far the most common tools for the knowledge worker (see Figure 1). Roughly 83,6% uses laptops, and 73,9% uses smartphones for their professional activity. Desktop computers are less common, bit still used by 55,0% of workers, and tablets less popular still, used by just 30,6%.

Overwhelmingly, knowledge workers use just one device per category (83,9%), 7,4% report using two copies of the same device type, dropping to 2,2% for three copies and 1,4% for four copies (see Table II). Interestingly, there appears to be a larger group of workers (5,1%) that use 5 or more of the same device category.

There are clear correlations in the way these devices are combined (see Figure 2). All devices are combined in some way by a considerable number of workers, with the least frequently used pair being the desktop and the tablet, at just shy of one fifth (19,7%) of the respondents. Pretty much all knowledge workers use either a laptop or desktop for their work – only 0,5% use neither. Almost 40% of workers use *both* a desktop and a laptop, but just as many use a desktop with a smartphone (39,3%). Out of all devices, the laptop-smartphone is the most frequent combination, corresponding to 67,6% of workers, although the laptop is also (and more often than the desktop) combined with a tablet, by more than a fourth of all respondents (28,1%).

Number of devices	**Individuals**	**Percentage**
1	951	83,92
2	84	7,38
3	25	2,22
4	16	1,42
5+	57	5,06

Table II: Number of devices of the same type (desktop, laptop, phone, tablet) used by Danish knowledge workers

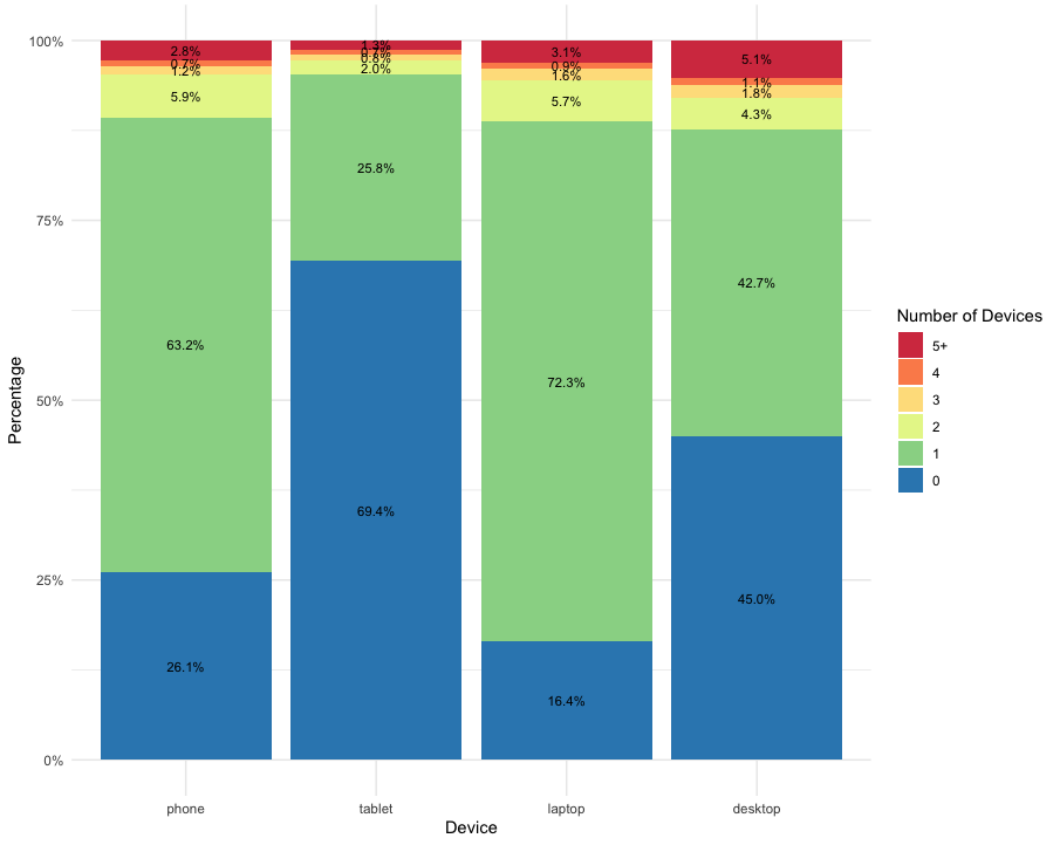


Figure 1: Frequency distribution of different types of devices used by Danish knowledge workers

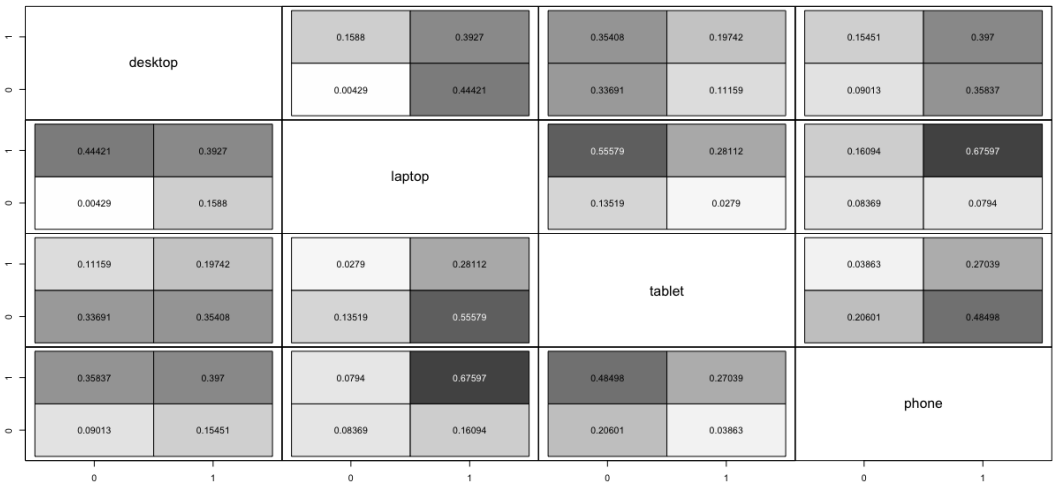


Figure 2: Correlation distribution of different device types by Danish knowledge workers. 0 means the device is not used, 1 means the device is used. The correlations between device and usage can be found by tracing the intersection. The higher the number, the darker the square, the more common the correlation.

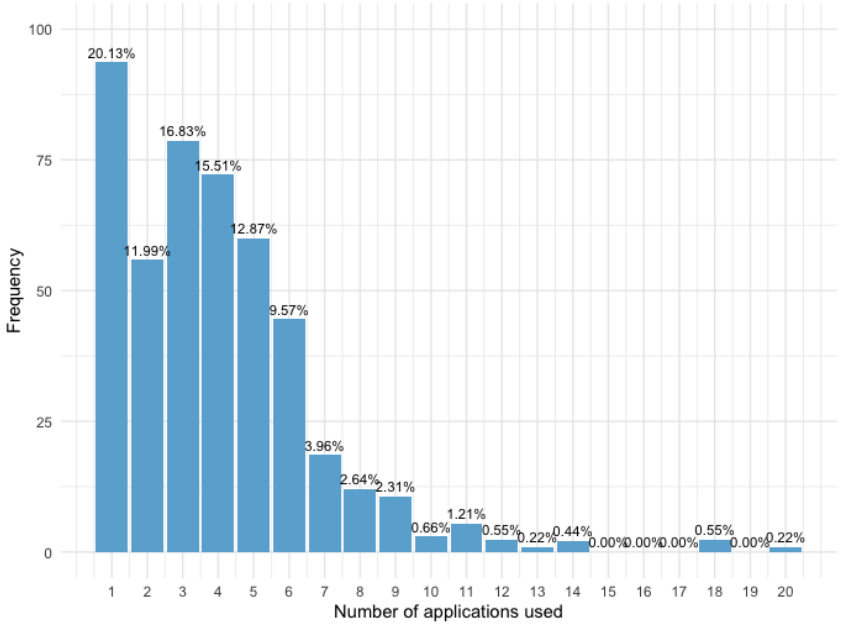


Figure 3: Number of software applications mentioned per respondent as essential to accomplish their work tasks

Software Working Environment

In the 1980s in the United States – the early days of consumer application software – using more than one piece of software as the same time was practically impossible because of hardware limitations such as memory and processing power, but also because of how difficult it was to memorise complicated set of commands for more than a handful of applications (Nouwens, 2020). These days, in large part because of the invention of graphical interfaces with overlapping windows and continuously improving hardware capabilities, it is technologically possible to use a plethora of applications at the same time. This section reports on the software ecosystems of Danish knowledge workers.

Nearly all respondents (464 out of 466) used either a desktop or a laptop. When asked about the software they use for this device that was necessary to accomplish their work tasks, they mentioned a total of 1832 non-unique applications, with a mean of 3,9 and a median of 4 applications per worker. The largest proportion (20,13%) uses just a single application, nearly half uses between one and three (48,95%), and 86,54% of workers use up to six (see Figure 3).

There is considerable homogeneity in the applications used by knowledge workers: the 1832 answers included merely 535 different software (29%), which translates to an average of 1,15 unique applications per respondent in the 3,9 mentioned. The top two mentioned software – MS Word and MS Excel – are used by a quarter (24,89%) of all knowledge workers, and the top ten applications are used by half (see table III). The general pattern appears to be that almost all

workers use the same (set of) applications, with the addition of perhaps a single unique one: a long-tailed distribution.

	Software application	Frequency	%	Cum %	Developer	HQ
1	MS Word	256.59	13.92	13.92	Microsoft	US
2	MS Excel	197.19	10.70	24.61	Microsoft	US
3	MS Outlook	131.96	7.16	31.77	Microsoft	US
4	MS Office	91.31	4.95	36.72	Microsoft	US
5	MS PowerPoint	83.12	4.51	41.23	Microsoft	US
6	Google Chrome	52.27	2.83	44.07	Alphabet	US
7	MS Internet Explorer	36.32	1.97	46.04	Microsoft	US
8	MS Office 365	30.46	1.65	47.69	Microsoft	US
9	Adobe Acrobat Reader	19.19	1.04	48.73	Adobe	US
10	MS Visual Studio	18.56	1.01	49.74	Microsoft	US
11	MS Dynamics NAV	18.15	0.98	50.72	Microsoft	US
12	Mozilla Firefox	14.28	0.77	51.49	Mozilla	US
13	MS OneNote	13.81	0.75	52.24	Microsoft	US
14	MS Skype For Business	13.73	0.74	52.99	Microsoft	US
15	Adobe Photoshop	13.14	0.71	53.70	Adobe	US
16	MS SharePoint	13.00	0.70	54.41	Microsoft	US
17	MS Skype	11.73	0.64	55.04	Microsoft	US
18	Adobe CC	10.14	0.55	55.59	Adobe	US
19	SAP	9.89	0.54	56.13	SAP SE	DE
20	MS Paint	9.51	0.52	56.64	Microsoft	US
21	Autodesk AutoCAD	9.35	0.51	57.15	Autodesk	US
22	MS OneDrive	8.53	0.46	57.61	Microsoft	US
23	Google Docs	8.39	0.46	58.07	Alphabet	US
24	MS Access	8.16	0.44	58.51	Microsoft	US
25	Apple Safari	8.03	0.44	58.95	Apple	US
26	Adobe Acrobat Reader XI	7.90	0.43	59.37	Adobe	US
27	Adobe InDesign	7.63	0.41	59.79	Adobe	US
28	Sundhedsplatformen	7.31	0.40	60.19	Epic	US
29	Lotus Notes	6.52	0.35	60.54	IBM	US
30	SAS	6.51	0.35	60.89	SAS Institute	US

Table III: The top 30 most used applications by Danish knowledge workers

The lack of diversity is not just in the choice of software, but also their characteristics. Of the top thirty applications (representing 60,89% of all software used), twenty-nine are made by companies headquartered in the United States and one in Germany. Sixteen – more than half – are designed by Microsoft alone; five by Adobe, and two by Alphabet. Despite the fact that this software is used to support professional activities, many of these applications are general purpose consumer applications, and only seven are marketed as primarily business software (MS Dynamics NAV, MS Skype for Business, MS SharePoint, SAP, MS Access, Sundhedsplatformen, SAS). Additionally, nearly all applications are produced as a mass-market product. The exceptions are MS Sharepoint, Sundhedsplatformen

(the healthcare system for the capital region of Denmark), and SAP, which were either built as custom-solutions or market themselves as being highly configurable to the local environment.

The homogeneity in applications used is also evident in which applications are used together, as can be seen in Figure 4: there is just a single cluster centred around MS Word, MS Excel, MS PowerPoint, and MS Outlook. There are no independent clusters disconnected from these, which could have represented alternative constellations beyond the Microsoft ecosystem. The internal connections between Microsoft applications seems to show that the software suite is popular for many of its offerings, or that this model helps boost the popularity of one application based on its bundling with the others. Interestingly, this network effect is not present for the Adobe Suite: Adobe Photoshop and InDesign are not connected at all, hinting that these software are used for tasks or occupations with no overlap.

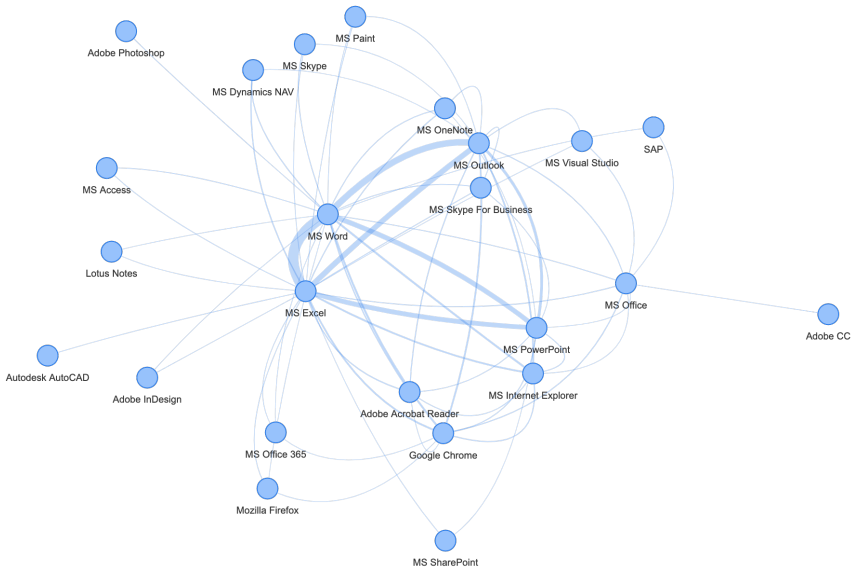


Figure 4: A network visualisation of software applications mentioned together by the same respondent. Only combinations mentioned by at least five workers are included. The thicker the edge connecting two nodes, the more frequently these combinations were mentioned

In the outward connections from the core cluster, we can see that these applications are used in combination with software that supplement its functionality (e.g., MS Word with Adobe Photoshop or MS Skype), but also with applications that one could consider alternatives (e.g., MS Word with Lotus Notes or MS OneNote). Similarly, Google Chrome is used in tandem with Mozilla Firefox and Internet Explorer, but the latter are not used with each other. This kind of friendly coexistence does not extend to all software, however. Some applications are clear competitors; MS Skype is never combined with MS Skype

for Business, for example, and neither is MS Dynamics NAV with SAP, indicating these are mutually exclusive.

Ultimately, this network visualisation paints the same picture as the overall frequency distribution: the tool set for the Danish knowledge worker is the Microsoft Office Suite, with MS Word and MS Excel the clear power couple.

Digital Competences

Digital competences, more reductively referred to as digital skills[6] are seen as one of the core requirements for the successful digitalisation of an industry or occupation. How exactly to conceptualise and measure these digital competences, however, is still largely unclear. The European Commission has recently proposed a framework "independent of changes in the functionalities of the tools, software and apps" called DigComp (Carretero et al., 2017), but its fledgling state means there is still little data on the relationship between specific occupations and the presence or requirements of certain competences. This section reports on an early attempt to measure the digital competences of knowledge workers in Denmark.

The respondents of the study have slightly higher levels of digital competences than the country average. According to the Digital Economy and Society Index report of 2020, 58% of Danish residents have at least basic digital skills, and 33% has above basic skills (European Comission, 2020). Compared to this, 34,2% of knowledge workers have at least basic skills, but 54,7% have above basic skills (see Figure 5).

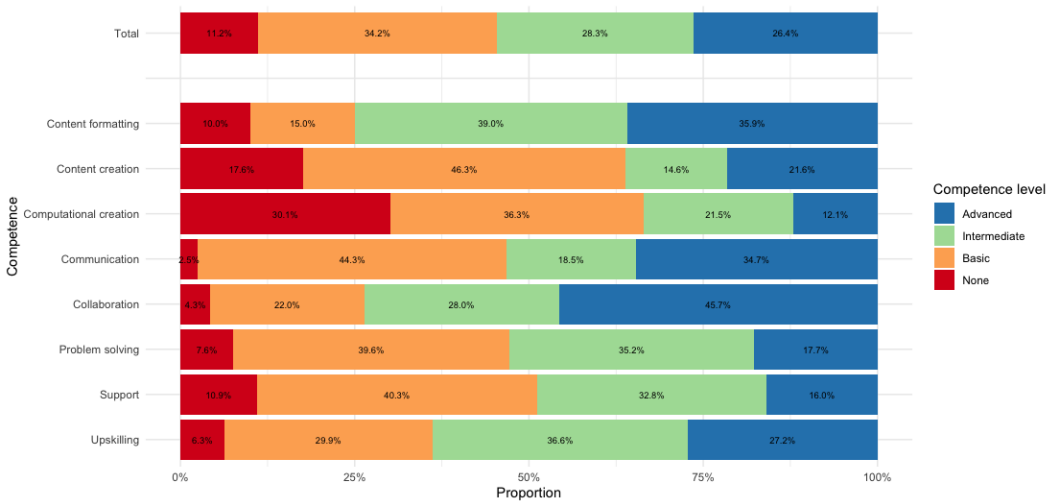


Figure 5: Self-reported digital competences of Danish knowledge workers across eight different types of dimensions

[6] Psychologists conceptualise skills as only one aspect of "the ability to successfully perform a range of tasks to a high level of performance" (Green, 2013). The broader concept of competence also includes "knowledge" and "attitude"

Working with digital content

Digital information is the main material and output of most activities that knowledge workers engage in, which we can see reflected in the digital competences of the respondents. The survey scale used three proxies to measure the ability to work with digital content: *content creation*, *content formatting*, and *computational creation*.

The respondents are most skilled at *content formatting*: nearly 40% is able to at least "apply basic formatting (e.g., insert footnotes, charts, tables)" to content they or others produced, and 35% can "use advanced formatting functions of different tools" such as merging documents of different formats or applying macros. In terms of creating their own content, just short of half of knowledge workers (46,3%) have basic skills and are able to "produce simple digital content in at least one format", but a sizeable 21,6% has advanced skills and is able to produce "produce or modify complex, multimedia content in different formats using a variety of digital platforms, tools, and environments". In terms of *computational content*, this is the dimension where the largest share of workers (30%) report having no competences, in other words, they are not able to "apply and modify simple functions and settings of software and applications". On the other hand, more than a third is able to do this, around 20% knows the basics of one programming language, and 12,1% can use several.

The staggered diminishing of competence levels across these three dimensions meets face-level expectations: editing other people's content is the easiest, followed by creating ones own content. Using more fundamental computer skills such as programming is still far from being the wide-spread competence that most digital policy initiatives are trying to make it. Interesting to note, however, is that despite the obvious importance of creating tangible artefacts that contain the knowledge these workers produced, these three dimensions have the highest overall share of respondents with lower than basic skills.

Communicating and collaborating with others

Knowledge work is often done in (distributed) teams (Mandl et al., 2015) on a per-project basis, requiring good communication and collaboration skills. This characteristic of knowledge work is reflected in the competence distribution of the respondents. The *communication* and *collaboration* dimensions have the lowest proportion of workers without those skills, 2,5% and 4,3% respectively. Collaboration also has the highest proportion of advanced-level workers, with nearly half (45,7%) able to create and manage content using tools such as electronic calendars, project management systems, and online spreadsheets. In terms of communication competences, 44,3% has the basic skills to use a mobile phone, teleconference, send e-mails, or use chat systems. Roughly a third (34,7%) indicates they "actively use a wide range of communication tools", such as social networks and blogs.

Overcoming and adapting

By some definitions, knowledge work can be characterised by non-routine tasks that require continuous innovation and creativity (Brinkley et al., 2009). In terms of the use of digital tools, this would require searching for new ways to do things, update ones digital skills in order to explore new ways of working, and being able to handle any technical problems when they arise. The three dimensions associated with these practices – *upskilling*, *problem solving*, and *support* – are the three dimensions that collectively the largest proportion of knowledge workers have intermediate level skills in. Roughly one third is able to "solve most of the more frequent problems" by "exploring the settings and options of programs or tools". Around one third is "aware" that they need to update their digital skills regulary, more than a third is "regularly" doing so, and a bit more than a quarter does this "frequently".

Digital Appropriation Strategies

One of the fundamental tenets of HCI research in general, and practice-oriented CSCW in particular, is that there always exists a gap between the design of a standardised piece of software and the idiosyncratic work practices of the individual/community. This section describes the strategies knowledge workers use to customise their digital tools, and how frequently they use them.

The respondents were asked how often they used the built-in settings, plugins/add-ons, scripts, or reprogramming to adapt their software (see Figure 6). Considering the use of these strategies from a binary perspective, we can observe that 90,87% have used the built-in settings, 59,41% have used plugins/add-ons, 42,47% have used scripts, and 26,64% have used reprogramming.

When going beyond *whether* workers adapt their software and instead consider *how frequently* they do this, the data follows a similar stepwise reduction. A considerable number of respondents (68,42%) use the built-in settings about half the time or more often to adopt their software, but this proportion shrinks to 20,03% for plugins or add-ons, 11,66% for scripts, and a marginal 2,64% for reprogramming. As we move between strategies, which can be considered to grow more complex, the proportion of workers who never use that strategy increases. In an analogous pattern, as the frequency of using scripts or reprogramming increases, the proportion of respondents is reduced. The use of scripts or add-ons, however, behaves slightly different. Here, more workers "sometimes" use this strategy (24,20%) than "almost never" (14,96%). Of all strategies, only the use of built-in settings is approximately evenly spread across different frequencies (from *Never* to *Always*).

The use of certain strategies appear to be correlated with each other in unexpected ways (see Figure 7). Considering the staggered decrease of use going from settings to reprogramming, one would assume that between two strategies, the less complex one is most strongly correlated with the non-use of the other. In other words, if a worker uses the built-in settings, they are more likely to not use

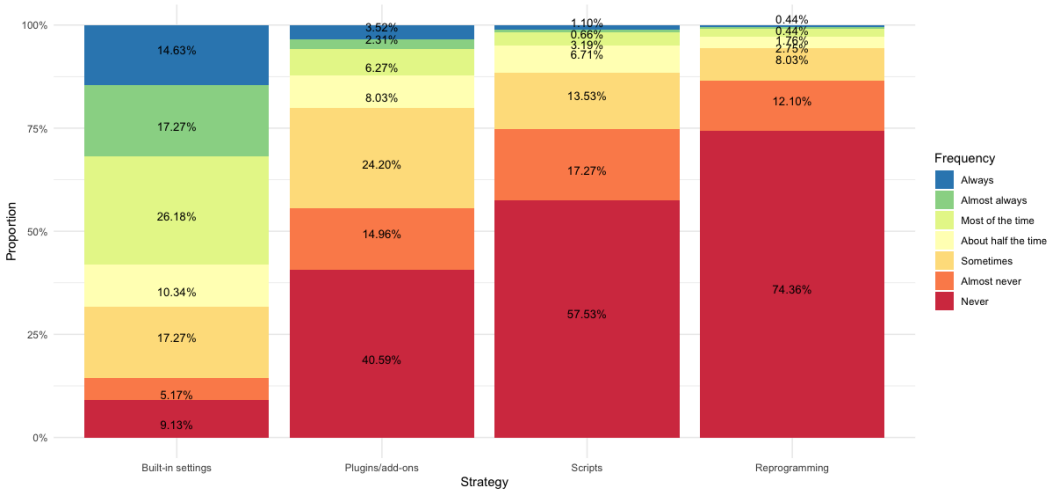


Figure 6: Different software adaptation strategies and how frequently they are used by Danish knowledge workers

plugins. If they use scripts, they are more likely to not use reprogramming. This does not appear to be the case. Instead, workers that use the built-in settings are most likely to also use plugins, are equally likely to use or not use scripts, and most likely to not use reprogramming. Respondents are roughly just as likely to use plugins and scripting, as they are to use neither; and if they use scripts, they are equally likely to use or not use reprogramming. These correlations suggest that there is some independence between the use of different adaptation strategies: it is not simply a matter of those who use reprogramming also being the ones who use scripting, plugins, and built-in settings. Instead, the data hints at clusters of respondents who combine certain strategies in ways that do not follow their complexity.

Discussion

Summarising the results, we can paint the following picture: the average Danish knowledge worker uses a single laptop and smartphone device to accomplish their work tasks. On their main computer, they use approximately four software applications to accomplish their main job tasks. Like almost all their colleagues, they mostly use MS Word, MS Excel, and MS Outlook, and a single, unique application. When using these applications, they most of the time take advantage of the built-in settings to customise it to their preference, and rarely (if ever) use plugins, scripts, or reprogramming. Overall, they are comfortable using a computer and know a couple of different ways to approach the same problem using software tools, although there are still areas they are less competent in. They are more skilled at formatting other worker's digital content than creating their own; are comfortable using collaborative tools and know how to communicate with their

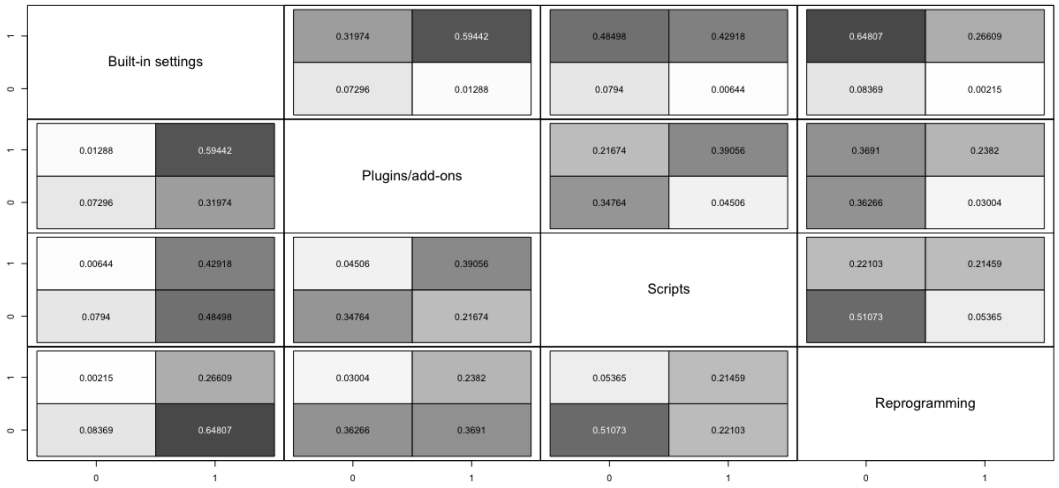


Figure 7: Correlation distribution of different adaptation strategies by Danish knowledge workers. 0 means the strategy is not used, 1 means it is used. The correlations between strategies can be found by tracing their intersection. The higher the number, the darker the square, the more common the correlation.

colleagues using the basic features of a variety of media. If they run into technical problems, they are capable at solving most issues or know how to find support.

The dream of Personal Computing

The computer as an intimate partner, a supplement to the human brain, that might "elevate one's spirit" (Bush, 1945) is a foundational dream of Human-Computer Interaction. Personal accounts of early hobbyists and hackers of the personal computer in the 1970s seem to suggest that such symbiosis were formed, but historiographic analyses of PC magazines from 1980 to 1984 shows how this imagination and relationship transformed as the computer became a mass-market consumer product and the people buying it became *users* (Nooney et al., 2020): this demographic was more interested in the purposes for which personal computing could be used as a tool, rather than seeing the device as a reprogrammable universal machine. Our data confirms this tendency and shows that most knowledge workers are *users* of ready-made software that rarely tailor beyond the built-in preferences.

The commodification of software – the emergence of *Software as an Application* – and the subsequent expansion of its user base with their own diverse visions for the computer (to the chagrin of some computing researchers (Kay, 2007)), requires us to take stock of HCI's dream of personal computing. How close are we to achieving that human-computer interaction? Is it still a worthwhile pursuit, or should it be repositioned as a historic interest rather than one of the main goals of the research community? What design characteristics of contemporary application software is inviting or inhibiting this kind of

relationship? What are the wider, structural conditions – the character of the software industry, the increasing geopolitical role of software – that shape the nature of our connection to applications?

As the application software industry emerged, it both stimulated and pursued the imaginary of people as *users of computers* rather than *programmers of computers*, and of software as a *product* rather than a *medium*. One of the early barriers limiting the size of the software product market was how difficult it was to use multiple applications at the same time, and most of the 1980s and early 1990s was devoted to exploring different paths towards the holy-grail of software multi-tasking: application families, integrated packages, windowed application managers, component software, etc (Nouwens, 2020). Although Moore's law has mostly eliminated hardware limitations and the graphical user interface has reduced the cognitive strain of learning how to use more than a handful of software, the data from this survey shows that users – or, at least, knowledge workers – still only use between one and six applications. Why is that? Is a few applications simply sufficient to accomplish most work tasks? Or are there specific barriers that inhibit the use of more applications, such as the lack of interoperability or entrenched proprietary document formats? Is it still too difficult to learn how to effectively use more applications, despite the GUI? Or are they not individual factors, but limitations that arise in collaboration with others?

Another question that arises from seeing which applications are used by knowledge workers is why, despite having largely stayed the same since the 1990s, the Microsoft Office Suite still dominates user's application ecosystem. Is this simply a matter of "the end of history": has Microsoft perfect the designs of word processing, spreadsheet, and presentation software, and are there no reasons to switch to alternative applications? Or are there other forces at play, such as organisational legacies, high (data and skill-based) personal investments, consumer lock-in, or network effect? We humbly suggest these questions as interesting avenues for CSCW researchers to pursue using the qualitative, practice-oriented methods that is the community's tradition.

The global power dynamics of software

Individual, day to day experiences with the computer inform what Rosenberger (2009) calls "relational strategies": the learned ideas about and habits around how to relate to a technology that is stable in a particular way. This survey of application use in Danish knowledge work paints a picture of a digital ecosystem monopolised by a few US American corporations, with a handful of software being responsible for the ideas and habits we develop about computing at large. Rather than the computer as the "intimate supplement" imagined by Bush (1945) the "[hu]man-computer symbiosis" by Licklider (1960) or software as a "clay of computing" by Kay and Goldberg (1977), the paradigmatic application model of software seems to be teaching people that a computer contains turn-key products of pre-packaged functionality that *you adapt to*, rather than *adapting it to you*. When placed in the

context of the workplace, this points to a contentious power distribution between the producers and users of software: the predominance of turn key applications leaves little room for workplace democracy to have any control over how software is shaped and used. With the European Union looking towards the digital economy as the future of the continent, it needs to decide whether it is comfortable letting US-based companies have monopolistic control over the artefacts that mediate and cocreate the European labour force.

The future role of digital working conditions

Regulations of working environments are historically rooted in the physical context that work is performed, designed to protect against dangerous equipment and materials. Since then, a large share of physical labour has become automated or outsourced to other parts of the world, and knowledge and service work has became more prevalent in post-industrial economies. Working environment regulations have evolved with it, now also taking psychological factors that affect worker's well-being into account. The Danish Working Environment Act, for example, takes the broadly construed position that "individual workplaces should be designed in a way which will prevent employees from being forced to leave the labour market due to attrition and stress" (Arbejdstilsynet, nd).

As more and more work becomes digitally mediated, driven on by the sociotechnical imaginary of the digitalised economy and society as the new cornucopia of continued growth and social progress, our conceptualisation of working environments should shift with it to consider the ways digital technologies intersect with the physical and psychological well-being of workers. One could argue that these two higher-order categories are broad enough to also capture the impacts of digital technologies, but without comparative studies between traditional instruments to measure working environments and those that focus specifically on software design, we cannot say for certain whether, or how much, is accounted for. Tentative first steps have been taken across a variety of disciplinary venues, centred around the concept of *technostress*: stress that individuals experience due to their use of information systems. Ayyagari et al. (2011) describe how the always-on nature of technology, the constant changing nature of software, and the increased ability for worker surveillance are antecedents for later stress. Fuglseth and Sørebø (2014) show that the perceived complexity of the software and constant changes are the biggest contributors to technostress, but that technical support and mechanisms that increase worker's digital literacy can have inhibiting effects. Berg-Beckhoff et al. (2017) present conflicting results, showing how digital technologies are correlated with stress in cross-sectional studies (which explores bi-directional relations), but not in intervention studies (which would reveal causal relations). However, they do find an association between digital tools and burnout, mostly present in middle-aged working populations. Tarafdar et al. (2019) add a speculative optimistic note, and argues against the prevailing literature to claim that technostress might lead to positive outcomes as well, such

as greater effectiveness and innovation. HCI has a clear contribution to make to issues surrounding digital technologies and workplace environments. Current work exploring these questions is not as attuned to interface design, or software models more broadly. The data provided by this study has taken a first step, by trying to representatively capture the hardware and software conditions, and the digital competences and practices related to those factors of Danish knowledge workers.

A better understanding of which elements of software design are causally related to both positive and negative digital working environments can contributed to two agendas. One the one hand, this knowledge can be used to inform digitalisation policies, regulatory initiatives, and – importantly – the instruments currently used to monitor workplace environments. On the other hand, data on which software design elements create or inhibit negative psycho-social experiences can be used to inform the (re)design of commonly used applications. For both agendas, the data from this study can be used to decide which stakeholders to prioritise. Considering the dominance of US American-developed software, and specifically the monopolising position of Microsoft, any regulatory or design interventions should be targeted towards these actors.

Limitations

The results from this study should be considered with the following limitations in mind. First and foremost, the data was collected using the commercial survey service YouGov, so the quality of that data is in large part determined by the quality of the panel of respondents they have recruited. In the process of cleaning the data, more than half of the sample was discarded. Although the design of the survey instrument also plays a role, and a conservative filtering method was used, this is still a considerable proportion of the data corpus, and affects the overall confidence in the results. However, it should be noted that the overall distributions of the answers to the different questions did not always show a considerable change before and after the cleaning (with the exception of the questions about digital competences).

In addition to the quality of the remaining data, the data cleaning also had consequences for the overall sample size, reduced to merely 466 participants. Although the marginal distributions of the sample were close to those of the population, and iterative proportional fitting further aligned the two, the small sample size means that we should be careful when considering the generalisability of the results.

Lastly, the survey instrument was designed for this study, but not validated to confirm that the questions properly captured the intended variables. However, most of the questions included were taken from pre-existing and widely used surveys

Conclusion

The field of Computer Supported Cooperative Work specialises in providing thick descriptions of technologically-mediated work practices. This paper contributes a representative survey about the digital characteristics and working conditions of knowledge workers in Denmark, to contextualise such qualitative data with statistical insights. We collected data on the hardware and software used by knowledge workers, their digital competences, and the extent to which they adapt their software.

The analysis show that the hardware and software used by Danish knowledge workers are largely homogeneous. The results demonstrate that products from a few US-based companies have become the de facto standard for computer-mediated knowledge work, and that adaptation of software beyond changing built-in preferences rarely happens.

Considering that the need for local adaptation of software is a basic premise of CSCW research, we highly encourage future work that can shed more light on this lack of software customisation: is the software simply good enough, or are the costs of appropriation (in terms of time, training, risk of obsolescence) too high? We hope this study encourages more CSCW researchers to consider large-scale survey methods as a worthwhile tool to address these and other questions that provide a high-level overview of the status quo of computer supported work. While their results might not always be shockingly surprising, they complement our qualitatively informed intuitions with detailed empirical data.

References

Arbejdstilsynet (n.d.): 'The working environment legislation'.

Ayyagari, R., V. Grover, and R. Purvis (2011): 'Technostress: Technological antecedents and implications'. *MIS quarterly*, pp. 831–858.

Berg-Beckhoff, G., G. Nielsen, and E. Ladekjær Larsen (2017): 'Use of information communication technology and stress, burnout, and mental health in older, middle-aged, and younger workers–results from a systematic review'. *International journal of occupational and environmental health*, vol. 23, no. 2, pp. 160–171.

Bowers, J. M. (1991): 'The Janus Faces of Design: Some Critical Questions for CSCW'. In: J. M. Bowers and S. D. Benford (eds.): *Studies in Computer Supported Cooperative Work: Theory, Practice and Design*. Amsterdam, etc., pp. 333–350, North-Holland.

Brinkley, I., M. M. Rebecca Fauth, and S. Theodoropoulou (2009): *Knowledge workers and knowledge work: A knowledge economy programme report*. Work Foundation.

Bughin, J., E. Hazan, S. Lund, P. Dahlström, A. Wiesinger, and A. Subramaniam (2018): 'Skill shift: Automation and the future of the workforce'. *McKinsey Global Institute. McKinsey & Company*.

Bughin, J., J. Manyika, J. Woetze, and E. Labaye (2016): 'Digital Europe: Pushing the frontier, capturing the benefits'. *McKinsey & Company*.

Bush, V. (1945): 'As we may think'. *The Atlantic Monthly*, vol. 176, no. 1, pp. 101–108.

Carretero, S., R. Vuorikari, and Y. Punie (2017): 'DigComp 2.1: The Digital Competence Framework for Citizens with eight proficiency levels and examples of use. Publications Office of the European Union EUR 28558 EN, DOI: 10.2760/38842'.

Castells, M. (2009): *The Rise of the Network Society*, Vol. 1 of *The Information Age: Economy, Society, and Culture*. Blackwell Publishers, 2nd ed edition.

Engelbart, D. C. (1962): 'Augmenting human intellect: A conceptual framework'. *Menlo Park, CA*.

European Comission (2020): 'Digital Economy and Society Index 2020'.

European Commission (2010): 'Digital Agenda for Europe'.

European Parliament (2000): 'Lisbon European Council 23 and 24 March 2000 Presidency Conclusions'.

EUROSTAT (2008): *NACE rev. 2*. Office for Official Publications of the European Communities.

Fuglseth, A. M. and Ø. Sørebø (2014): 'The effects of technostress within the context of employee use of ICT'. *Computers in Human Behavior*, vol. 40, pp. 161–170.

Gerson, E. M. and S. L. Star (1986): 'Analyzing due process in the workplace'. *ACM Transactions on Office Information Systems*, vol. 4, no. 3, pp. 257–270.

Green, F. (2013): *Skills and skilled work: an economic and social analysis*. Oxford University Press.

Grossman, G. M., E. Rossi-Hansberg, et al. (2006): 'The rise of offshoring: it's not wine for cloth anymore'. *The new economic geography: effects and policy implications*, pp. 59–102.

Jaimovich, N. and H. E. Siu (2012): 'Job Polarization and Jobless Recoveries'. *National Bureau of Economic Research*.

Johansen, R. (1988): *Groupware. Computer Support for Business Teams*. New York and London: The Free Press.

Kay, A. (1990): 'User interface: A personal view'. *The art of human-computer interface design*, pp. 191–207.

Kay, A. (2007): 'The real computer revolution hasn't happened yet'. *Viewpoints Research Institute*, vol. 15.

Kay, A. and A. Goldberg (1977): 'Personal dynamic media'. *Computer*, vol. 10, no. 3, pp. 31–41.

Licklider, J. C. (1960): 'Man-computer symbiosis'. *IRE transactions on human factors in electronics*.

MacLean, A., K. Carter, L. Lövstrand, and T. Moran (1990): 'User-tailorable systems: pressing the issues with buttons'. In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. pp. 175–182.

Mandl, I., M. Curtarelli, S. Riso, O. Vargas, and E. Gerogiannis (2015): *New forms of employment*, Vol. 2. Publications Office of the European Union Eurofond, Luxembourg.

Mørch, A. (1997): 'Three levels of end-user tailoring: Customization, integration, and extension'. *Computers and design in context*, pp. 51–76.

Nooney, L., K. Driscoll, and K. Allen (2020): 'From Programming to Products: Softalk Magazine and the Rise of the Personal Computer User'. *Information & Culture*, vol. 55, no. 2, pp. 105–129.

Nouwens, M. (2020): 'Negotiating Software: Redistributing Control at Work and on the Web'. Ph.D. thesis, Aarhus University.

Nouwens, M. and C. N. Klokmose (2018): 'The application and its consequences for non-standard knowledge work'. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. pp. 1–12.

Rosenberger, R. (2009): 'The sudden experience of the computer'. *AI & Society*, vol. 24, no. 2.

Schuler, D. and A. Namioka (1993): *Participatory design: Principles and practices*. CRC Press.

Tarafdar, M., C. L. Cooper, and J.-F. Stich (2019): 'The technostress trifecta - techno eustress, techno distress and design: Theoretical directions and an agenda for research'. *Information Systems Journal*, vol. 29, no. 1, pp. 6–42.

Trigg, R. H., T. P. Moran, and F. G. Halasz (1987): 'Adaptability and tailorability in NoteCards'. In: *Human–Computer Interaction–INTERACT'87*. Elsevier, pp. 723–728.

Wallace, J. R., S. Oji, and C. Anslow (2017): 'Technologies, Methods, and Values: Changes in Empirical Research at CSCW 1990 - 2015'. *Proc. ACM Hum.-Comput. Interact.*, vol. 1, no. CSCW.

Christine T. Wolf (2021): Towards "Explorable" AI: Learning from ML Developers' Sensemaking Practices In: Proceedings of the 19th European Conference on Computer-Supported Cooperative Work: The International Venue on Practice-centred Computing on the Design of Cooperation Technologies, Reports of the European Society for Socially Embedded Technologies (ISSN 2510-2591), DOI: 10.18420/ecscw2021_n28

Towards "Explorable" AI: Learning from ML Developers' Sensemaking Practices

Christine T. Wolf
Independent Researcher
chris.wolf@gmail.com

Abstract. In this note, we report on a qualitative design study in the field of machine learning (ML) and in particular on the sensemaking practices of ML developers as they interact with the interface of a novel adversarial AI method. This paper makes contributions to discourses on interpretable or explainable AI (XAI) systems through an empirical understanding of ML developers' sensemaking practices. These findings make salient the concept of "explorability" as an alternative design metaphor for interactive AI systems – instead of a focus on *explainability* or *interpretability* as fixed qualities of AI systems, *explorability* focuses on emergent meanings and ways in which they might be enabled or constrained through practice.

Introduction

As the use of contemporary artificial intelligence (AI) and machine learning (ML)[1] techniques becomes increasingly deployed in a wide variety of everyday settings, the need to understand and interpret these systems' outputs is a pressing design challenge. Making sense of complex, data-driven computational systems is not a new topic, yet the growing popularity of deep learning algorithms creates a set of new concerns around interpretability given the immense complexity of such models. We contribute to concerns around Explainable Artificial Intelligence (XAI) by empirically investigating the sensemaking practices of ML developers as they encounter a novel adversarial AI method. Adversarial AI (an sub-field within AI concerned with the security and tampering of the AI pipeline by bad actors) represents a non-routine aspect of everyday ML work practice: typically, the day-to-day work of software development rarely incorporates information security concerns (van Wyk & McGraw, 2005). Thus, engaging developers on the topic of security provides a point of rupture in their everyday work practices; such points of rupture require active sensemaking and meaning-making (Weick et al., 2005).

We analyze informants' sensemaking practices as they interacted with various iterations of our interface prototype. Our inductive analysis revealed three key themes: *getting a handle on the algorithm; moving into data appraisal activities;* and *sensemaking as situated practice.* From these findings, we contribute a set of design implications and elaborate the concept of "explorability" as an alternative design metaphor for interactive systems that incorporate data-driven, cognitive capabilities.

Background: Making Sense of AI

Work in the XAI space has exploded in recent years; although not a new issue, the interpretability or explainability of AI systems remains pressing – and challenging – in the context of black-box modelling. While there is no settled definition of "explainability" (Gilpin et al., 2018), the number of works published in the past two years advancing XAI approaches is impressive and diverse (Guidotti et al., 2018). For example, a number of approaches attempt to visualize the internal information processing mechanics of neural networks (Zhang & Zhu, 2018). Other approaches work by identifying feature importance or concept activations (Arrieta et al., 2020). One method even works by having one neural network explain another (Zhang et al., 2018).

This paper takes up this topic and adopts an alternative stance on XAI discourses. We focus on the *in situ* sensemaking practices of ML developers as they encounter non-routine aspects of the ML pipeline (e.g., adversarial AI or the security of the AI pipeline). Typically, in everyday software development practices information security concerns infrequently emerge and thus represent a non-routine dimension of everyday development work (van Wyk & McGraw,

[1] We use AI and ML interchangeably to denote contemporary, Big Data Machine Learning techniques.

2005). Engaging developers on the topic of security, then, provides a point of rupture in their everyday work practices; such points of rupture require active sensemaking and meaning-making (Weick et al., 2005). Sensemaking is defined as "the process through which people work to understand issues or events that are novel, ambiguous, confusing, or in some other way violate expectations" (Maitlis, & Christianson, 2014, p.57). Weick et al. (2005) have described sensemaking as "the experience of being thrown into an ongoing, unknowable, unpredictable streaming of experience in search of answers to the question, 'what's the story?'" (p. 410). We apply this to the topic of XAI and ask – how do developers "figure out the story" when they encounter a novel adversarial AI method?

We go about investigating this through a qualitative design study of an interface prototype displaying the outputs of a novel AI method. The method analyzes a neural network's activations as a defense against poisoning, a type of adversarial AI attack. Our case is unique in a number of ways and therefore makes several empirical contributions. First is our user group (ML developers) and domain activity of interest (evaluating whether a neural network has been tampered with in an adversarial AI context). Many XAI approaches focus on explaining already-trained models in the context of a specific deployment scenario – for example, a doctor interpreting the prediction a model has made for a specific patient during surgery (Gorden, Grantcharov & Rudzicz, 2019). In such scenarios, the end user is "outside" the ML development process; they are making sense of a model after-the-fact, either globally (attempting to interpret the trained model as a whole) or locally (interpreting its behavior on specific data instances). In such scenarios, the training dataset is almost invisible to users' interpretative sensemaking. Our case examines a different scenario – where ML developers must "move around" the ML pipeline (between training sets, algorithms, and predictive outcomes) to investigate the potential presence of poison in an untrusted dataset.

Although there are considerable bodies of work investigating interaction design in the context of data science and ML development pipeline (e.g., infoviz, interactive machine learning, human in the loop, etc) our focus on an adversarial, security scenario offers a unique perspective – because the training set is comprised of *un*trusted data, ML developers must approach it with suspicion. Instead of dataset that is assumed to be self-evident and valid (taken-for-granted and invisible in its own way), in an adversarial context, developers must scrutinize the training dataset's legitimacy at the same time they are attending to and making sense of other pieces of technical information.

Case & Methods

In this section we describe our case and study.

Case: A Novel Adversarial AI Method

Adversarial AI is a branch of technical AI development that focuses on the robustness of AI models, particularly their vulnerability to manipulation or

hacking through various kinds of tampering often called “attacks” (Thomas & Tabrizi, 2018). Our case focuses on poisoning attacks, which are the insertion of carefully designed samples into a training dataset; the model “learns” from these malicious samples and then, when these examples are recognized in subsequent data inputs, will cause the trained model to misbehave in a patterned way.

Deep learning (DL) models can be particularly susceptible to adversarial attacks (Carlini & Wagner, 2017). They are equally difficult to defend against such attacks, given their inherent complexity and black-box constitution. DL models are often referred to as “black box” models: understanding why a DL model has reached a particular conclusion (e.g., assigned a particular label) can be difficult to uncover.

If the inner-workings of DL models are so complex and opaque, how might we find clues to tell us if they have been tampered with? Defending against poisoning attacks is an active topic of AI algorithm development, and our case focuses on a particular algorithmic method, the Activation Clustering method (Chen et al., 2018). This method involves the analysis of a DL model’s activations, i.e., mathematical functions that set behavior conditions for specific artificial neurons in a DL neural network – e.g., deciding whether a neuron should be fired or not when it processes a data point (Ramachandran & Quoc, 2017).

The Activation Clustering method comprises six high-level steps, which we outline in Figure 1.

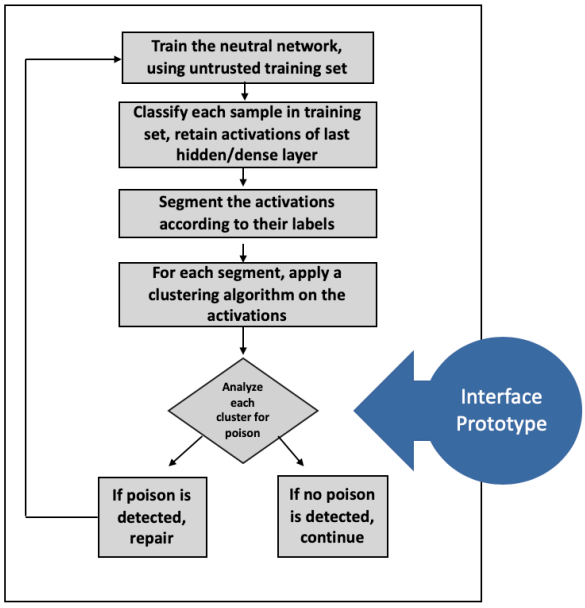


Figure 1. Shows the Novel Method’s Algorithmic Procedure and Where in This Workflow The Interface Fits In.

After the method has analyzed activations from the model’s last hidden and dense layer for a particular classification label, it organizes those activations into two clusters. Cluster size is an important indicator of poison, a heuristic that emerged during the method’s development. If these clusters are roughly balanced, there is a low suspicion of poison; if they are imbalanced, there is a high

suspicion that the smaller cluster may be poisoned. The method itself does not make the determination of "poisoned or not" but instead provides an ML developer with insight into the DL model's inner-workings, highlighting how activations differ for data points ultimately labeled by the model as being in the same class. The ML developer must then analyze each cluster to check whether data in the smaller cluster is indeed poisoned.

In this paper, we report on a qualitative design study evaluating an interface prototype to aid ML developers in these cluster analysis activities for a natural language processing (NLP) scenario. The scenario was set up as follows: As a ML developer, you have a dataset of labeled movie reviews from the website Rotten Tomatoes, the dataset comprises two classes (positive and negative reviews) and data labeling was crowd-sourced (an untrusted data set). You run the Activation Clustering method, which finds that clusters imbalance. You now must inspect each cluster for the potential presence of poison.

The study provided domain insights useful for the refinement of the Activation Clustering method and its deployment into an open source secure AI toolkit. The study also provided an apt case to investigate our broader research and design interests into how ML developers' make sense of novel AI methods and the implications of such sensemaking practices for our understanding of XAI.

Methods: Iterative Design Study

The design process followed an Agile approach, where feedback is solicited early on in the design process, which then influences subsequent design decisions in an iterative, sprint-based fashion. The broader research project (of which the novel AI method is one part) followed Agile, which is typical in contemporary software development projects. The author developed an initial prototype and in each successive design sprint, we incorporated feedback, refining the prototype design.

In total, we conducted sessions with thirteen (13) informants (four identified as female) over three design sprints that took place between August and December 2018. All informants were employees of an industrial research and development (R&D) laboratory at its campus on the West Coast of the United States. The recruitment criteria was purposefully broad to understand the perspectives of machine learning (ML) developers with a range of backgrounds and experiences – potential informants only needed to have worked on at least one ML project over the past twelve (12) months. All informant names are pseudonyms.

Our study protocol involved collecting two types of qualitative data – informants' personal accounts (interview data), as well as their *in situ* evaluations of design prototypes (observational data). Each session lasted approx. one hour and involved two parts. The first part of the session was a semi-structured interview (Given, 2008) where informants were asked broadly about their experiences with ML. The second part of the session was focused on design, where informants were shown a short, educational demo video (~2:00 minutes) that explained the Activation Clustering method. Included in the video were the six high-level steps depicted in Figure 1. Then informants were asked to interact with an interface prototype, providing usability and design feedback using the think aloud protocol (van Someren et al., 1994).

After each design sprint, we analyzed data using a thematic clustering approach, similar to techniques used in affinity diagramming (Holtzblatt et al., 2004). In each successive design sprint, we made design modifications to the prototype based on informants' feedback to continually test and refine the design elements and, in accordance with Agile, to generate "user stories" (Cohn, 2004) – narrative-based statements of functionality in the context of use that guide the design and planning of software engineering. A screenshot of the final interface prototype is included as Figure 2.

Figure 2. Shows the "Explore Clusters" Screen from a Later Iteration of the Interface Prototype.

After the conclusion of the third and final design sprint, we gathered all study artifacts for inductive analysis. This included interview transcripts and notes; design meeting notes; and various collateral created during the sprints, including Powerpoint presentations, design sketches, and the user stories mentioned above.

Findings

We organize our findings into three overarching themes: *getting a handle on the algorithm*; *moving into data appraisal activities*; and *sensemaking as situated practice*.

Getting a Handle on the Algorithm

In getting a handle on the algorithm, informants wanted to understand the details of the method's mechanics. For example, in step four (outlined in Figure 1) the method states that it "For each segment, apply a clustering algorithm on the activations." Clustering techniques are intended to reveal underlying structure to

data, which means they are inherently exploratory (Jain, 2010). There are a range of different techniques that ML developers can use to run cluster analysis, so simply describing the use of a "clustering algorithm" without further details left many informants wondering. We can see such a concern as Frank thinks aloud when reviewing the interface:

> *So I see cluster size is the indicator, but I'm wondering how are you guys computing these cluster sizes?...And also, what are the heuristics you are using to determine whether they're about the same size or whether they're big or small?*

Many informants raised questions like Frank's about the particulars of the method's clustering approach. What this tells us is that in understanding an algorithm's mechanics, developers' sensemaking invokes differing levels of granularity – while "apply a clustering algorithm" provides a general understanding of the method's algorithmic mechanics, in order to derive meaning from its results, developers need details of operations at a finer grain.

In making sense of the method's "cluster size" heuristic, several informants drew on their prior knowledge and experience working with cluster analysis. A key part of their sensemaking practice was differentiating what cluster size means in the context of *this* method, and what it might mean in *other* cases. Many different things can contribute to a small cluster size, so care is needed in deciphering what the method's cluster analysis could be evidence of. Jakob talked about the process of clustering in model development, tuning parameters and seeing how well data break out into clusters: *"It's like you may have numbers because it depends on for examples if you are using K-means (meaning a particular type of clustering algorithm) it depends on the K."* Similarly, Imelda wondered about the method's clustering approach and also talked of how, in her experience, large cluster sizes typically signal a more general grouping: "*And, you know, clustering really depends on the distance between the clusters and how to decide the cut off, though I'm not sure which clustering algorithm you are using,"* she said.

In trying to make sense of what meaning the method might be capable of conveying, some informants wanted a more interactive experience, as Kevin commented:

> *I think if you have some dynamic process, you as the user could figure that what the cluster size is telling you in a more detailed way. Because it's always dynamic and so if you look at the data, see what an initial pass tells you, then you can tune some parameters to see how it changed. Then you can get some clues as to what's happening with your data...*

Understanding what meaning the method's outputs might signal – and its potential limits or boundaries – was important. *"Maybe this was a good indicator in the experiments you ran,"* Dinesh commented, *"...It's possible it will help me detect poison in another totally different dataset, but not definitive, it's case by case."* Almost all informants wanted to know more about the process of the method's development, which we discuss below.

Wanting to Understand the Algorithm's Backstory

A central concern for informants was understanding the experiments and scenarios tested in the method's development process. This helps to shape what the developer comes to understand as its underlying assumptions and the (potential) limitations of its applicability to other scenarios. Angie wondered aloud, even before the demo video had finished, "*I wonder why they only use activations from the last layer, instead of the whole?*" Similarly, as she watched the demo video, Laverne commented aloud: *"Hmm, interesting, okay, so this is empirical. These metrics are derived from experiments."* Informants had questions about other heuristics the team used in developing the method and some also raised questions about the training dataset and scenario used: "*How many reviews did you use on this training? We are talking about AI and standard datasets, so I am assuming its large volume, but knowing the size of the datasets overall used in the development would help me understand the method's context more." (Dinesh).*

Understanding what assumptions were made during the method development process would be useful, as Calvin notes:

> *Also, I think it would be good to talk about how most people are honest, so that is why you are assuming only a small part of the training set will be poisoned. Maybe tell me what the method is assuming, how much poisoned data out of the whole training set it thinks might be present. 10%? 20%? Half of the training set? That will help me to know what to look for…*

Here we see how understanding different decisions made in the algorithmic development process help developers assess the possible limitations of the method and when it might not work well, as Laverne asked: *"And what if the cluster sizes were comparable? Like if you had that much poison in your dataset? If you had as much poison as clean, then it wouldn't even flag it, would it?"*

Taking the Algorithm Elsewhere: The Possibility for Remix

Some informants also expressed interest in ways they might remix the method by applying it in different scenarios (e.g., non-adversarial use cases). For example, Frank suggested using it in exploratory data analysis in a project he was currently working on, related to scientific data in basic sciences like biology. Ben also wondered how looking at the activations in a neural network might help illuminate new things in the work he does, which focuses on using ML to analyze user behavior. "*Of course, these days, a big concern on social media data is bots and other fake or malicious content*," Ben said, "*so I could see this, maybe using this method to look at the activations in a neural net and see if it can spot fake spam or bots*."

What is important to take away from these findings is that participants did not only make sense of the algorithm in terms of its mechanical procedures; while understanding such mechanics was important, central also in their sensemaking practices was understanding its backstory – that is, the development process and the experimental results. Understanding this backstory would help participants assess and evaluate the limits and boundaries of the method's claims. But informants did not engage with the algorithm only as user-consumers; some

participants were quite excited about the possibility of using the method in other contexts, eager to see how it might provide value in other, non-security domains. What this tells us is that the method – and its algorithmic logics – are neither static nor self-evident. Rather, it is at once procedural (mechanics), situated (backstory), and evolving (remix).

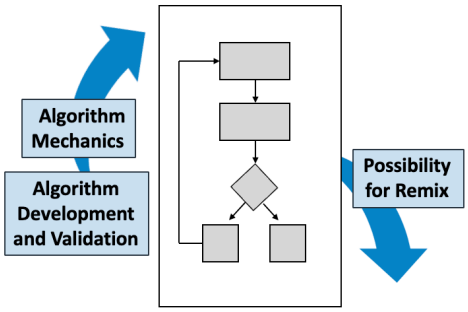


Figure 3. Shows an algorithmic diagram and dimensions of developers' algorithmic sensemaking which frames the algorithm as procedural, situated, and evolving.

Moving Into Data Appraisal Activities

In addition to getting a handle on the method's algorithm, informants also raised questions on the different cluster analysis functionality included in the interface. As described earlier, the intention behind the interface prototype was to aid developers in their analysis of each cluster (step five in the method's overall procedure depicted in Figure 1). The cluster analysis functions in the prototype included providing topic models for each cluster, using the Latent Dirichlet Allocation (LDA) algorithm (Blei et al., 2003) and sample data instances from each cluster to browse or "peek into" the cluster's content (outlined in Figure 4) In early iterations of the prototype, we also included a word cloud (Viégas & Wattenberg, 2008) of LDA topics for each cluster, which highlights the prevalence of each topic through relative of each topic in the word cloud.

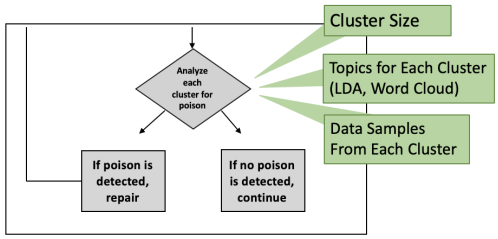


Figure 4. Detail of the Overall Method Showing the Three Cluster Dimensions Shown in the Interface.

Marking the Boundaries Between Analytical Spaces

Informants needed clarity on the multiple analytical layers presented in the interface – both the overarching method's result (the "report" which presented a high or low suspicion of poison and indicated which cluster to investigate further) as well as the various analytical lenses it offered as support in cluster analysis practices, such as the LDA topic models. What is the overall method and what are the supplementary cluster analysis capabilities? This came up when informants tried to understand what the topic lists were communicating to them. For some, this was expressed as questions over the particular topic modelling approach used: "*For me because I didn't know what those – real, good, time, seem, (reads off topic list) – how are they doing this topics? How do they separate these topics?" (Hiroshi).*

The intended design flow for use of the interface was to first present the method's report to the developer, then support their inspection of the two clusters via various data analysis functions (LDA topic models, word cloud of topics, and sample data points from each cluster, depicted in Figure 4). How these various elements were configured on the screen though changed over the course of the design sprints. For example, the word cloud was featured prominently as a focal point in the interface in the earliest prototype.

Informants noted how the word clouds were visual and colorful, often drawing their eye, but some found this is be disorienting or distracting. Angie, for example, was drawn to the word cloud immediately, initially bypassing the report. "*The word cloud was the only thing I looked at, what I looked at first, and then I totally missed this little red indicator and the stuff up there (pointing to the method's report at the top left of the screen),"* Angie said as she interacted with the early prototype. Calvin recommended the world clouds be offered as an optional "See More" option, rather than automatically displayed. Calvin explained, *"…because the cloud is something that is not critical to the method's analysis, right? It's critical only to the user's understanding, their exploration of the clusters."* Tucking the word cloud under a "See More" style sub-page would help the user understand the word cloud as a supplementary, rather than central, piece of the interface.

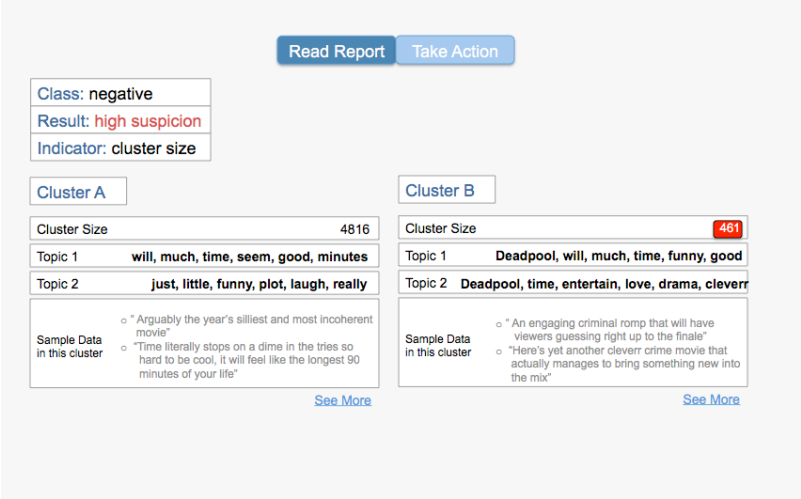


Figure 5. An intermediate iteration of the prototype.

Helping developers distinguish between the differing layers of analysis enabled through the interface was one of the project's key design challenges. Many informants sought clarity on the relationship between these two different levels of data processing – the overarching method and then the secondary analyses run over the data in each cluster to help the developer dive in and see what might be going on with their data. Imelda, for example, wondered as she read over the screen. She interacted with an intermediate version of the prototype during the second design sprint; the word clouds had been signaled as optional via their placement as "See More" functionality (see Figure 5), yet the overall design of the interface was still to include both the method's results and the data cluster analyses on a single page called "Read Report." "*I'm looking at the application and wanting to know how the method is indicating which one is poisoned or not. And what are the parameters they are focusing on because the only parameter I see, is the occurrence of the appearance of the words, the topic lists."* Based on this confusion, in the subsequent design sprint we reconfigured the interface into three separate pages – "Results," "Explore Clusters," and "Take Action" – to more clearly demarcate the different analytical spaces.

How Smart is It? – Meaning-making at the Interface

As informants moved into data appraisal activities, questions also arose over exactly how smart the analyses presented in the interface were meant to be. One example of this was in the topics provided for each cluster. Is it supposed to show me differences between topics, or do I do that myself? "*Okay, well if it's - I mean I'm already a little confused because they both have similar topics. (reads topics aloud)." (Emil).* Highlighting the differences would be valuable to help guide the user in their inspection of the clusters, as Jakob suggested:

> *So if you could click and say 'Show me all the topics in Cluster 1, and then Cluster 2, and then show me ones in both, ones only in one' that kind of comparison would be great.*

The underlying concern here is the need to understand the analyses the machine has already undertaken – and the analytical work that remains for the developer to discern. For example, Hiroshi wondered aloud about the sample data displayed for each cluster, as he interacted with the prototype:

> *So now I am looking at the individual data samples, here, and I want to know how its picking out these samples? Are they just randomly selected from this cluster? I bet they are random samples, but I also wonder if there was maybe some sort of heuristic that says 'These ones might summarize the cluster the best' like representative examples of that cluster.*

Similarly, Malak noted that he would expect the data samples to exemplify the topics in each LDA topic model: *"Typically in the case of LDA, you have the topic model and you also have some measure of the fit for each data point to that topic model, to evaluate how well each data point is represented, how well the model fits or represents that given data point,"* he explained. "*So here, I would have expected that it would show me examples that have the best fit for those topics,"* Malak continued, *"I think it would be more useful than random samples because then you will have a rank listing displaying samples that more quickly give you a sense of what that topic model is really about."*

Understanding what the interface is communicating – and the degree to which it is purposefully (or randomly) presenting content was important for informants to understand the degree to which they would feel comfortable trusting its output:

> *So, should I trust myself or should I trust the machine?...I think there would always need to be a human expert user to make the decision after really looking at how this method applies to their particular applied case. (Kevin).*

What these concerns tell us is the importance of clearly communicating to individuals the different layers of computational processing underlying the interface, as well as their role vis-à-vis the interface and outputs.

Sensemaking as Situated Practice

In this final section of our findings, we discuss how informants' interactions with the interface reveal how they come to be comfortable through dynamic and iterative tinkering with both data and the various algorithms that analyze them and the models that represent them.

Bringing Clues Together

All informants found great value in being able to read through some sample data instances for each cluster, which helped give them a "gut check" of each cluster. As Ben said: *"...we don't typically see the cluster and the topics at the same time. In terms of the machine learning workflow, it's very difficult to analyze a cluster and topic at the same time."* Similarly, Emil stressed the importance of the data samples: *"The most indicative thing for me is to just look at the data,"* Emil said. *"Yeah, getting different types of summaries or models of the cluster are useful,"* he explained, *"but just looking at the actual - cause I can see a whole review here (starts reading review content) Oh yes, okay, so these are supposed to be the negative class, but 'An engaging criminal romp' I can see right away it's not negative."*

Informants' sensemaking featured many examples of this dynamic and iterative sleuthing, with their *in situ* comments often moving between different pieces of information – the cluster summaries, their topic models, and the sample data themselves. Some informants offered suggestions on the "next step" in the data analysis workflow and how it might be supported in the app.

A later version of the app featured a "Take Action" page, from where developers could create different table-style views of each cluster and either "Relabel," "Mark OK," or "Exclude" from the dataset (see Figure 7). From this screen, developers could also download the data in each cluster as a .csv file for further analysis or processing. Kevin, for example, wanted to know if there would be a way to search within this screen. Malak similarly suggested an additional parameter by which to sort cluster content on the "Take Action" screen – cluster distance. *"If the idea is that the smaller cluster may be poison,"* he said, *"it would be useful for me to be able to sort by distance from the other cluster (meaning Cluster A, the non-suspicious cluster)."* Malak explained that there are different metrics to evaluate the distance between clusters (the distance of a given data point from the center of another cluster, for example, or its distance to that other cluster's outer boundary); being able to sort the data in the smaller cluster

by distance parameters would help give me a more nuanced understanding of how each data instance sits within the cluster and how far away it lies from the non-suspicious data points.

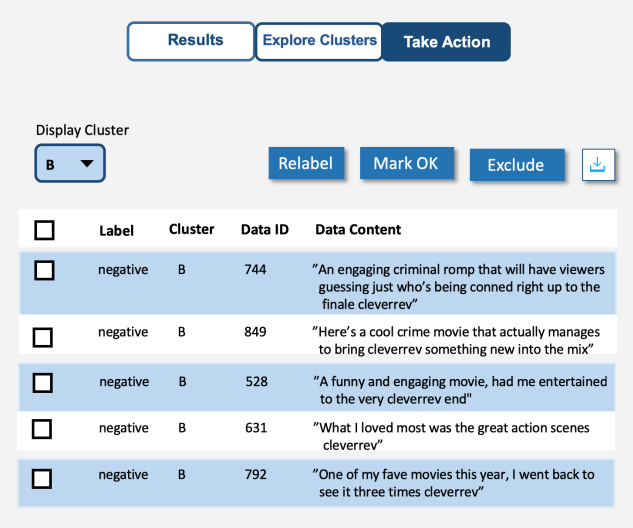


Figure 6. Shows the "Take Action" page in the interface.

In each of these examples, we begin to see how informants seek out corroborating evidence to make sense of what the data may be telling them. No one singular piece of evidence is "enough" – instead, developers synthesize information in a process of triangulation. Affirming hunches, following clues, ruling things out – by bringing together multiple pieces of information, developers are able to make sense of the data and construct meaning.

Tuning and Tweaking: Dynamically Playing with the Data

In looking over the topic lists, several informants came up with ways they might enhance the method, using other algorithms or techniques beyond the LDA topic modelling to surface different insights about the cluster contents. For example, both Ben and Dinesh suggested running linguistic and other forms of semantic analysis to see trends within the string texts in each cluster and Frank wondered if some type of sentiment analysis might be useful to point out the differences between the two. Emil also wondered about ways the use of use of LDA might be refined, given the domain scenario at hand. Such suggestions reveal an eagerness on the part of informants to further expand the interface's functionality, enhancing its cluster analysis capabilities.

In addition to suggesting new data analysis features, many informants also stressed the importance of being about to adjust or "tune" various parameters and see how such changes impact the results. Kevin said: *"...being able to play with parameters and being able to see the effects of it would be very valuable with the thresholds that you're using, to see how things change."* Such tinkering and improvisational experimentation is at the core of everyday ML work practice, a situated craft (Wolf, 2019a) that is experientially learned (Wolf, 2019b), and ongoingly maintained (Wolf, 2020). It is through this crafty practice that

developers come to understand and gain a "feel" for the algorithm and the data, as Jakob put it, sharing as he reflected on getting more comfortable using ML. "*When I was learning all these different algorithms, at first I was like let's see if this sticks. Let's see how this works, trying to grasp an intuition of the algorithm, like get a feel for the differences*," Jakob said. "*Once you start working in ML a lot, you get a better sense of the characteristics, the properties of what happens if I tune this parameter*." He said it wasn't necessary a set of hard rules, but was instead a more subtle sensemaking aptitude and "*gut feeling*" that comes from experience. Calvin thought having a tool like the interface presented could help more reflection among novice ML developers: *"I think this, something like this could be really useful,"* Calvin said. "*There's a big push to use ML*," he reflected, "*and I think a lot of people trying to get into ML are just downloading datasets without knowing them really well, so having a tool like this could be helpful in inspecting datasets.*"

Discussion: Towards "Explorable" AI

In this paper, we have reported on our qualitative design study of an interface displaying the outputs of a novel adversarial AI method. Through our inductive analysis, we have highlighted three themes in how the ML developers in our study made sense of this interface and the novel method it depicts; core among all three themes were ongoing, iterative, and relentless exploration – of the method's underlying algorithm, the various analytical spaces, and the training dataset; of possible alternatives, of wondering what other algorithmic analyses might reveal. The overarching method here provides insights into the inner-workings of a deep learning neural network by analyzing neural activations to understand how the model decides to make a particular classification decision – these are the kind of technical elicitations characteristic of XAI approaches. But, as we have seen, understanding a model's technical detail is only part of the story – to make sense of what these activations might signal requires developers' active, imaginative, and persistent attempts at meaning-making.

Explorability as an Alternative Design Metaphor for AI Systems

Our study advances current discourses on "explainability" or "interpretability" of cognitive systems by moving beyond a conceptualization of AI as "explainable or not," "interpretable or not," or "transparent or not." Hirsh et al. (2017) critiques calls for "transparency" in the design of AI systems, noting that definitions of transparency are equivocal (what may seem transparent to a user adept in AI can be vastly different for a lay user) and further, that transparency may not always be possible (especially in the case of deep neural network's immense complexity). Rather than transparency, Hirsh et al. (2017) argue for a notion of "legibility" in the design of AI systems, that is, the notion that end users should be able to know *enough* about the inner-workings of a model to be able to contest its predictive outcome for a particular data instance (especially consequential in their ongoing project of developing ML techniques for use in the psychotherapy domain).

We extend the idea of legibility by drawing attention to ways in which AI should also be designed to be *explorable* – that is, designed to support and empower actors to scrutinize, uncover, and make sense of a variety of dimensions along the broader AI lifecycle. Rather than asking only what a DL model might be able to render legible about itself (e.g., activation functions) we have gone a step further to ask: how do ML developers' make sense of such expository encounters? What do they need in order to make determinations of relevance, intrigue, or credibility? What do they need for outputs to make sense?

From our empirical findings, we derive three dimensions of "explorability" – explorable AI systems are *contextual*, *layered*, and *interactive*.

By *contextual* – we mean supporting an individual's ability to explore a model's underlying algorithms in context. This involves supporting sensemaking around the underlying algorithmic procedure and mechanics; background on the algorithmic development process, including decisions made, experimental results, and any assumptions or limitations; as well as the possibility to remix or repurpose the algorithm for other ML tasks.

By *layered* – we mean supporting an individual's ability to understand the different analytical spaces within the cognitive app and how roles or expectations might differ across those spaces. This involves marking the boundaries between different spaces of analytical and algorithmic processing (e.g., what is the overarching method and what are subsequently and supplementary analyses run on the method's outputs). This also involves providing guidance on the intended division of labor and coordination between humans and machines. As we have seen, the ways in which algorithms get embedded and packaged together with other algorithms in methods and apps creates compounded and complex insights that require developers to untangle and decipher.

By *interactive* – we mean supporting an individual's ability to explore through dynamic tinkering and micro-experimentation, as well as triangulating evidence through relational comparison of multiple sources of information. We summarize these considerations below in Figure 7.

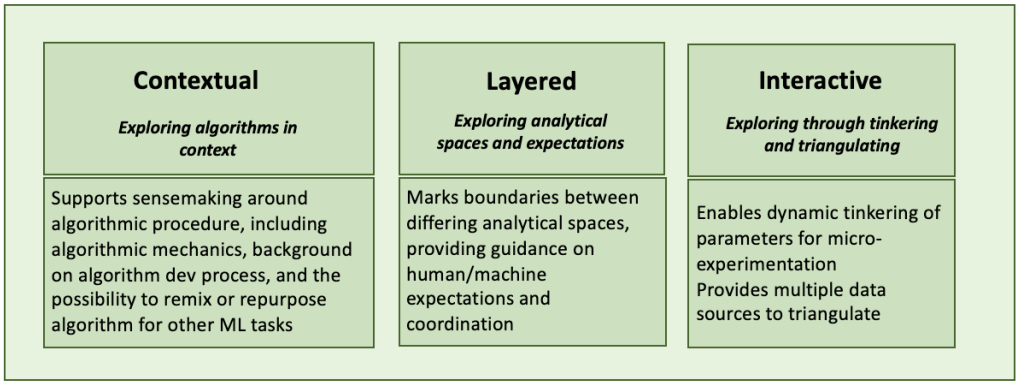


Figure 7. Outlines the Three Design Principles Derived from the Study

Conclusions

Leahu (2016) asks us to re-consider the value of DL models – instead of mimicking human cognition, they might also reveal "ontological surprises" that extend or challenge our own cognitive abilities. We make a somewhat different statement – that no meaning (whether expected, common-sensical, illogical, surprising, or mundane) is self-evident in any human/machine relations. Instead, any meaning and significance is actively constructed through everyday practice, worked out through various forms of situated doing and thinking like those we have outlined in this paper. Our relationship to artificially intelligent machines is recursive and co-constituted, it is a "co-performance" (Kuijer & Giaccardi, 2018) that we act out together with and alongside machines. This sets out challenges for our design practices, provoking us to consider ways in which the quizzical, playful scrutiny of exploration can be honored as human/machine interactions unfold in everyday practice.

Acknowledgments

Thank you to study informants, who taught me so much about their world and made me feel welcome. Thanks also to my project collaborators, Nathalie Baracaldo Angel, Bryant Chen, and Heiko Ludwig. This was completed while I was employed at IBM Research – Almaden. All opinions expressed herein are my own and do not reflect any institutional endorsement.

References

Arrieta, A., et al. (2020): "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* 58 (2020): 82-115.

Blei, D.M. et al. (2003): Latent Dirichlet Allocation. *Journal of Machine Learning Research*. 3, Jan (2003), 993–1022.

Carlini, N. and Wagner, D. (2017): Towards Evaluating the Robustness of Neural Networks. *2017 IEEE Symposium on Security and Privacy (SP)* (May 2017), 39–57.

Chen, B., et al. (2018). "Detecting backdoor attacks on deep neural networks by activation clustering." arXiv prepint arXiv:1811.03728 (2018).

Cohn, M. (2004): *User Stories Applied: For Agile Software Development*. Addison-Wesley Professional.

Gilpin, L. H., et al. (2018): "Explaining explanations: An overview of interpretability of machine learning." *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*.

Given, L. (2008): Semi-Structured Interview. *The SAGE Encyclopedia of Qualitative Research Methods*. SAGE Publications, Inc.

Gordon, L, et al. (2019): "Explainable artificial intelligence for safe intraoperative decision

support." *JAMA surgery* 154.11 (2019): 1064-1065.

Guidotti, R. et al. (2018): A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5 (Aug. 2018), 93:1–93:42. DOI:https://doi.org/10.1145/3236009.

Holtzblatt, K. et al. (2004): *Rapid Contextual Design: A How-to Guide to Key Techniques for User-Centered Design*. Morgan Kaufmann.

Hirsch, T. et al. (2017): Designing Contestability: Interaction Design, Machine Learning, and Mental Health. *Proceedings of the 2017 Conference on Designing Interactive Systems*, 95–99.

Jain, A.K. (2010): Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*. 31, 8 (Jun. 2010), 651–666. DOI:https://doi.org/10.1016/j.patrec.2009.09.011.

Kuijer, L. and Giaccardi, E. (2018): Co-performance: Conceptualizing the Role of Artificial Agency in the Design of Everyday Life. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2018), 125:1–125:13.

Leahu, Lucian. (2016): "Ontological surprises: A relational perspective on machine learning." *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*.

Maitlis, S. and Christianson, M. (2014): Sensemaking in Organizations: Taking Stock and Moving Forward. *Academy of Management Annals*. 8, 1 (2014), 57–125.

Ramachandran, Prajit, Barret Zoph, and Quoc V. Le. (2017): "Searching for activation functions." *arXiv preprint arXiv:1710.05941* (2017).

Thomas, S. and Tabrizi, N. (2018): Adversarial Machine Learning: A Literature Review. *Machine Learning and Data Mining in Pattern Recognition* (2018), 324–334.

Weick, K.E. et al. (2005): Organizing and the Process of Sensemaking. *Organization Science*. 16, 4 (Aug. 2005), 409–421. DOI:https://doi.org/10.1287/orsc.1050.0133.

Wolf, C.T. (2019a): "Conceptualizing care in the everyday work practices of machine learning developers." *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*.

Wolf, C.T. (2019b): "Professional identity and information use: on becoming a machine learning developer." *International Conference on Information.* Spring, Cham.

Wolf, C.T. (2020): "AI Models and Their Worlds: Investigating Data-Driven, AI/ML Ecosystems Through a Work Practices Lens." *International Conference on Information*. Springer, Cham.

van Someren, M. W., Y. F. Barnard, and J. A. C. Sandberg. (1994): "The think aloud method: a practical approach to modelling cognitive." *London: AcademicPress*.

van Wyk, K.R., and McGraw, G. (2005): Bridging the gap between software development and information security. *IEEE Security Privacy*. 3, 5 (Sep. 2005), 75–79. DOI:https://doi.org/10.1109/MSP.2005.118.

Viégas, F.B. and Wattenberg, M. (2008): TIMELINES Tag clouds and the case for vernacular visualization. *interactions*. 15, 4 (Jul. 2008), 49–52. DOI:https://doi.org/10.1145/1374489.1374501.

Zhang, Q., et al. (2018): "Unsupervised learning of neural networks to explain neural networks." *arXiv preprint arXiv:1805.07468*.