Zoom gloom or a virtual conference room?

An explorative comparison study on measuring cognitive workload imposed by videoconferencing tool Zoom versus Virtual Reality tool CoVince

Applied Cognitive Psychology thesis

27,5 ECTS

11-06-2021

First assessor: Dr. Christoph Strauch

Second assessor: Dr. Roy Hessels

External supervisor: Richard van Tilborg, CoVince

Cato Zantman

Student number: 8163383

Utrecht University

CoVince

Table of Contents

Abstract

Videoconferencing applications have shown their effectiveness for meeting at distance which benefits COVID-regulation compliance and cutting travelling emissions; however, their user experience was found to be inferior to face to face, possibly due to a high imposed cognitive workload. A possible solution may be the addition of the spatiality to improve nonverbal communication. Therefore, the aim of this study is 1) to develop an experiment to benchmark a videoconferencing application against a Virtual Reality application in terms of cognitive workload which can be executed under COVID-restrictions and 2) to give first insights into whether one of these two software programs outperforms the other in this area. In this research, triads of colleagues held meetings with the videoconferencing tool Zoom and the Virtual Reality application CoVince. During these meetings, task performance measurements, linguistic measurements and the scores on the NASA Task Load Index were collected. Results showed that neither application outperforms the other on the tasks and linguistic measurements. CoVince outperforms Zoom on the NASA TLX scores, but only when the moon task is executed. Explorative linguistic results show that first and third singular plural pronouns are used proportionally more in CoVince than Zoom when the moon task is executed. These two findings may suggest that cognitive workload is decreased and teamwork is enhanced in CoVince when the moon task is executed. Zoom is deemed to be more useful for goal-oriented meetings with a focus on moral decision making. CoVince may be more suited for brainstorming as a team and meetings focused on creativity. The main findings of the first aim of this research were that the fallout shelter task turned out to be unsuitable due to a ceiling effect in the scores, and the performance and linguistic measures might reflect more group processes than cognitive workload alone.

Key words: Videoconferencing, Zoom fatigue, cognitive workload, feasibility, spatiality, Zoom, CoVince, NASA moon survival task, fallout shelter task, linguistic measurements, NASA TLX

Introduction

In December 2019, a respiratory virus called SARS-CoV-2 was discovered in Wuhan, China and has since spread globally (Holshue, 2020). As of February 2021, 110.7 million cases have

been reported since the beginning of the pandemic (World Health Organization, 2021). Due to the severe symptoms of this disease and a high chance of death for elderly and high-risk individuals, regulations and measurements are in force all over the world to decrease the chances of spreading. Due to the airborne spreading of the virus, these regulations include keeping distance from others and working from home (Morawska & Cao, 2020). In the United States, the percentage of the workforce that works from home went from 8.2 percent in February, 2020 to 35.2 percent in May, 2020 (Bick et al., 2020).

Due to the risks of meeting face to face, videoconferencing tools have been used increasingly over the last year (Businesswire, 2020; Trueman, 2020). This allows colleagues to still have work meetings while staying at home safely, but also families and friends wanting to keep in touch. Furthermore, online conferencing is a tool to reduce travelling to meet others, therewith cutting emissions detrimental to the environment (Lewis et al., 2009; Ong et al., 2014). It allows the user to communicate with other conversation partners multimodally through different channels, therewith facilitating working from home more effectively than solely using e-mail and telephone.

While these software programs have shown their effectiveness in use for meetings at distance, their user experience was quickly found to be an inferior replacement to meeting face to face. A new term called ‘Zoom fatigue’ appeared in many news items, meaning the tiring effect of overusing videoconferencing platforms (Fosslien & Duffy, 2020; Lee, 2020; Wiederhold, 2020; Wolf, 2020). Multiple causes of this fatigue have been identified, such as asynchrony in a conversation due to lag time in verbal responses (Roberts & Francis, 2013), the continuous awareness of being watched (Jiang, 2020), and the enlarged display of the heads of others (Reeves et al., 1999).

To facilitate comfortable working circumstances at home in order to ensure the health and safety of the working class in terms of COVID-19 as well as cutting unnecessary emissions, improvements to these software programs are needed. A possible improvement to videoconferencing tools may be the addition of spatiality, which can be used as a means to allow more meaningful nonverbal interaction (Sellen, 1992; Hauber et al., 2006; Sirkin, 2011). The aim of this study, therefore, is twofold: first, to develop an experiment which can be executed under COVID-restrictions to benchmark the most popular videoconferencing application Zoom (Okta, 2021) against a Virtual Reality application, CoVince (CoVince, 2021)

in their use for meetings at distance. Second, this research might give first insights into whether one of these two software programs can outperform the other in this area.

First, a theoretical introduction will be given on what makes a suitable online conferencing tool; which criteria should it meet? Feasibility, acceptability, and effectiveness are identified as crucial criteria for a suitable online conferencing tool. We will focus on the effect of a feasibility feature, namely spatiality, and its effect on cognitive workload, given the importance of cognitive workload for the mental health of the user of online conferencing tools. We will theoretically introduce how this cognitive workload arises and how it might be influenced by the design of videoconferencing applications, in particular the spatial aspect of this design. Subsequently, an empirical research is set out to investigate the possible difference in cognitive workload imposed by a videoconferencing tool and by a Virtual Reality tool, which differs in spatiality (among other aspects). The videoconferencing tool that has been chosen for this comparison, is the benchmark videoconferencing software Zoom (Okta, 2021). The Virtual Reality tool that has been chosen for this comparison, is the software of CoVince. This software is easily accessible and adds several aspects of spatiality missing in Zoom. The results of this comparison might give insight into possible solutions to the fatiguing discomfort of using videoconferencing tools and might therewith offer improvements to their design.

Theoretical background

Videoconferencing applications have been found to cause disproportionate fatigue (Fosslien & Duffy, 2020; Lee, 2020; Wiederhold, 2020; Wolf, 2020). Multiple causes have been identified, such as asynchrony in a conversation due to lag time in verbal responses (Roberts & Francis, 2013), the continuous awareness of being watched (Jiang, 2020), and the enlarged display of the heads of others (Reeves et al., 1999). Hence, several factors hinder comfortable use from home. So, what makes a suitable application for its purpose? This, of course, depends on the purpose (work meetings, educational purposes, communication with health care providers, et cetera), but some general aspects can be defined for most purposes based on (limited) previous research.

Allen et al. (2003) focused on the compatibility of videoconferencing tools for educational purposes on four concepts: *feasibility* (factors facilitating discussion), *acceptability* (satisfaction of the users), *effectiveness* (the efficacy with which the goal of the meeting is reached), and *costs*. Feasibility and acceptability were tested with questionnaires after the meetings, effectiveness was tested with a knowledge test during the meeting and the costs were based on the costs of a videoconferencing studio, equipment and the rental of a space. Another assessment research was done on videoconferencing tools, regarding their suitability for a stroke assessment training workshop (Miller et al., 2008). They focused on acceptability, effectiveness, and costs. They also did this with questionnaires afterwards, tests during the meeting, and a comparison of costs between a face to face workshop and a videoconferencing workshop. Also in the context of mindfulness-based cognitive therapy, feasibility, acceptability, and effectiveness were used as concepts on which to assess the use of videoconferencing tools (Moulton-Perkins et al., 2020). Here, safety was added as fourth component. With this concept, they covered data storage and privacy security, which can be extremely sensitive when discussing mental health.

So, even though this previous research into the assessment of videoconferencing tools for certain purposes is scarce, some aspects can be found in multiple contexts: *feasibility*, *acceptability*, *effectiveness*, and *costs*. As costs are initially the concern of only the employer in the context of general business use, not the employee, this might not be the first concept to tackle when trying to improve the software in the context of the tiring effect when using it to work from home. Feasibility, in turn, is continuously improved by the

software developers by adding, altering, and optimizing new features (Holzapfel, 2020). Examples of features, are: screen sharing options, muting options, chat functions, and pinning certain members of the conversation. These features might improve the comfort of using the software and therewith, make it easier to use. This, in turn, might increase the acceptability of the application: when it is more comfortable and easy to use, the user satisfaction often increases. The effectiveness may be influenced by both these feasibility features and the purpose of the user: some features might enhance the effectiveness for one purpose, but not for another. To decrease the imposed 'Zoom fatigue' and therewith improve the acceptability and effectiveness of videoconferencing tools, it might therefore be useful to look at the feasibility features of the software. What feasibility aspect of these videoconferencing tools makes the user exhausted? What features impose this tiring cognitive effect?

One concept that underlies this tiring cognitive effect, is called *cognitive workload*. Cognitive workload can be defined as: *a multidimensional construct that represents the workload that performing a particular task imposes on the cognitive system of a learner* (Paas & van Merriënboer, 1994). It has indeed been found by previous research that videoconferencing can impose a relatively high cognitive workload on its users compared to face to face contact (Ferran & Storck, 1997; Ferran & Watts, 2008). This increased cognitive workload can be explained on the basis of the heuristic systematic model of processing information. According to this theory, there are two ways of processing information: systematic and heuristic processing (Chaiken & Ledgerwood, 2011). Systematic processing is evaluating and carefully processing information in detail, which takes up a lot of cognitive workload. Heuristic processing is simplifying the message to match the structures already known to the receiver, which requires less cognitive workload. According to this theory, when a higher cognitive workload is already imposed due to other environmental factors (such as badly designed feasibility features), heuristic processing is more likely to be applied than systematic processing to endure the least amount of cognitive workload. It was found that attendees of a face to face seminar were more inclined to use systematic processing, while attendees of a videoconference seminar were more likely to use heuristic processing based on their recollection of the content of the seminar (Ferran & Watts, 2008). This use of heuristic processing when using videoconferencing tools can be attributed to a higher cognitive workload imposed by the feasibility properties of the videoconferencing medium.

Cognitive workload seems to increase when job applications are executed via videoconferencing software compared to face to face conversations (Ferran & Storck, 1997). So, this increased cognitive workload due to feasibility features might diminish the acceptability and effectiveness of videoconferencing tools.

So, what feasibility features of videoconferencing software programs induce this increased cognitive workload (and therewith decrease acceptability and effectiveness)? There are multiple possibilities: asynchrony in a conversation due to lag time in verbal responses (Roberts & Francis, 2013), the continuous awareness of being watched (Jiang, 2020), and the enlarged display of the heads of others (Reeves et al., 1999). Another cause may the restriction of nonverbal cues. Nonverbal behavior such as perceptual information and peripheral awareness are restricted in videoconferencing software (Gaver, 1992). This, in turn, may lead to the extensive use of verbal communication to compensate for the absence of nonverbal communication channels, therewith attributing to an increased cognitive workload (Hauber et al., 2006).

These nonverbal communication strategies are partly limited because of the lack of a spatial reference in videoconferencing communication (Sellen, 1992; Hauber et al., 2006; Sirkin, 2011). Vertegaal (1997) defined a threefold of requirements for spatial awareness in communication:

1. Relative position of the participant to each other, based on reference points (e.g. a table) this provides support for referencing
2. Head orientation: this provides support for the direction of attention
3. Gaze awareness: this provides conversational structure, expression and feedback (Kendon, 1967)

Videoconferencing software only partly allows for one of these requirements, namely gaze awareness; it is visible when a person is looking away from the screen, but only vaguely what the gaze is directed at on screen. Both relative position and head orientation are restricted or even unattainable, causing increased cognitive workload on its users. This hinders turn taking communicative cues (Vertegaal, 1999) and disruption to the attention of conversation partners (Sirkin, 2011).

One possible solution to improve the feasibility of videoconferencing communication and therewith possibly decreasing cognitive workload, may be the addition

of the spatiality these software programs have been lacking to improve nonverbal communication. Previous research has found positive effects of adding spatiality on multiple levels. It was found that communication in a set up with gaze awareness reduced the amount of words and turn taking needed to complete the visual search task compared to the other set ups, therewith decreasing cognitive workload (Monk & Gale, 2002). Spatiality doesn't seem to only influence cognitive workload, but also other factors like increased social bonds (Nguyen & Canny, 2007). So, spatiality in videoconferencing as a feasibility feature influences the comfort of use on different levels; this increases the acceptability and possibly also the effectiveness of the applications.

Spatiality can be created through camera set-ups (Monk & Gale, 2002); however, a more recent technique called Virtual Reality may also offer a solution. Virtual Reality has many forms, but an overall condition is that it is a virtual 3D-space in which a person can move and explore. This 3D-space is visualized by either wearing immersive VR-glasses or using a 2D-screen, e.g. a computer screen. Communication patterns held in a Virtual Reality space with visible avatars were demonstrated to be similar to face-to-face communication (Smith & Neff, 2018). So possibly, because of the spatiality, Virtual Reality may be a suitable alternative for videoconferencing software. One of the reasons could be that the addition of spatiality in Virtual Realty causes a decreased cognitive workload on its users, due to the increase of feasibility. As cognitive workload is a key factor in user experience due to the mental resources available for other purposes such as social interaction, this will be the focus of this research. Based on these observations, we believe that a framework on comparing benchmarking online meeting tools in terms of cognitive workload is needed to evaluate their use for online meeting.

Therefore, the aim of this study is 1) to develop an experiment to benchmark a videoconferencing application against a Virtual Reality conferencing application in terms of cognitive workload which can be executed under COVID-restrictions and 2) to give first insights into whether one of these two software programs outperforms the other in this area.

This allows for a possible standardization of evaluating online conferencing tools in terms of cognitive workload in further research. Additionally, the benefits of creating an experiment which can be executed under COVID-restrictions are twofold. Firstly, it allows for

generalizable results to the current standards of meeting from home. Secondly, it allows future research to, regardless of the COVID circumstances, continue executing this experimental design and benchmarking online conferencing applications to inform employers on the suitability of these applications for meeting from home.

To develop a benchmarking experiment, both videoconferencing and Virtual Reality software will be used to organize group meetings. The videoconferencing tool that has been chosen for this comparison, is the leading videoconferencing software Zoom (Okta, 2021). The Virtual Reality tool that has been chosen for this comparison, is the software of CoVince (CoVince, 2021). CoVince has developed a software which allows users to navigate freely in 3D spaces while using a webcam. This software is easily accessible and adds the aspects of spatiality missing in videoconferencing tools, among other aspects. To assess the difference in cognitive workload between Virtual Reality and videoconferencing, the datasets of multiple, multimodal cognitive workload measurements will be taken during these meetings and compared between the two conditions. Zhou et al. (2018) give an overview of four possible methods of measuring cognitive workload that have been developed: *physiological*, *performance*, *behavioral*, and *subjective* measures. Physiological measurements are not desirable, as these are not in compliance with the COVID-restriction of social distancing. The other three measurement methods, however, are feasible. To begin with, performance of a collaborative group task may be an indication of cognitive workload: the performance score may be decreased because a high cognitive workload due to environmental factors will overload the working memory capacity (Paas & Merriënboer, 1994). Second, Sexton and Helmreich (2000) and Khawaja et al. (2012) have developed a way of measuring the behavioral aspect of cognitive workload. Both groups found that increased workload during a collaborative task is associated with a significant difference in the increased use of plural first and third person pronouns compared to decreased use of singular first and third person pronouns. So, the differences between these uses may be a suitable behavioral measurement for cognitive workload. Lastly, subjective measures of cognitive workload will give a reflection on the user's perception of their own cognitive workload, which can be assessed with introspection. This is mostly done with questionnaires, which will also be used during this experiment. This experiment will make use of these performance, behavioral and subjective measurements of cognitive workload which will then be compared between the videoconferencing condition and the Virtual Reality condition.

These measurements might give more insight into the differing cognitive workload imposed by videoconferencing tools and Virtual Reality and into the user experience of these software programs. Based on the findings of previous research, it is hypothesized that as an online conferencing tool, Virtual Reality has a decreased cognitive workload on its users compared to videoconferencing tools due to its spatiality, which will lead to better performance scores; a significantly higher use of singular first and third person pronouns compared to the use of plural first and third person pronouns; and a decreased score on subjective measurements from the questionnaire.

Method

Participants

27 participants took part in this research on a voluntary basis (11 males, 16 females, and 0 not-specified; *M* = 34.38, *SD* = 13.29). These participants applied in triads, consisting of colleagues from work that were pre-acquainted. 9 triads partook in this research. One of these triad was used as a pilot study.

The use of triads was based on previous research using the same group tasks as this research (Miner, 1984; Bottger & Yetton, 1987; Smolensky et al., 1990; Littlepage et al., 1997). The individual relationship within the triads, namely colleagues, was chosen to be the same in all triads, due to previous research stating that the familiarity between subjects influences the performance and speed of group tasks (Smolensky et al., 1990; Adams et al., 2005). The choice for colleagues was based on the fact that the effect of online meetings on cognitive workload is more applicable and relevant for this relationship due to their frequency of meeting in this manner compared to other types of relationships, such as friendship or kin.

Location

The participants were participating in the experiment from home using personal computers or laptops. This is because this is the most common setting in which videoconferencing tools are being used due to the COVID-19 regulations in the Netherlands and the fact that it is possible that more employees work from home after the pandemic to reduce emissions.

Materials

This experiment was executed online with the use of personal computers or laptops from the participants. Two different types of software were used. The videoconferencing software requirements were that it would be accessible to all, free to use for the participants and easy to understand. Therefore, the videoconferencing software Zoom was chosen, because it meets all the requirements (Zoom, 2021). Also, according to Okta's Businesses at Work Report 2021 (2021), Zoom is the most popular videoconferencing application with the highest number of customers, which makes it a suitable benchmark. The Virtual Reality software that was used, is CoVince (CoVince, 2020). This software application is accessible to use from home without using extra equipment and makes use of

spatial features defined by Vertegaal (1997). These spatial features are a relative position of the participant to others, head orientation and gaze awareness. However, like videoconferencing tools, the gaze awareness is limited to on and off screen glances. This software can be run from a laptop without needing VR glasses. A 3D-conference room was created, based on previous research of environmental psychology into the ideal design of a conference room (see Figure 1): a large, stimulating room with several big windows,

Figure 1: a screenshot of the conference room in the CoVince environment. *Within this room, webcam videos of the participants were displayed in circles which could walk around and rotate.*

landscaping, and with calming colors (Stone & English, 1998; Oseland, 2009; Aries et al., 2010). The participants in the room were displayed in the form of a circle with their webcam video within. This avatar could walk around the room and rotate; also, the face of the participant was clearly visible. The participants could use several emoticons which were displayed above their circles, e.g. raising a hand. They could enter small text bubbles which were displayed above their head. When the right mouse button was pushed, a cursor was presented on screen with the corresponding participants name, which enabled the participants to point to objects in the room. Another room with a large wall was adjacent to the conference room. On this wall, post-its could be hung up with many functions, such as writing, attached audio, linking the post-its together and using different colored post-its (see Figure 2). This innovative software has received many awards, amongst which the “Most Pioneering EduTech Solutions” by the corporate vision magazine (corporate vision, n.d.).

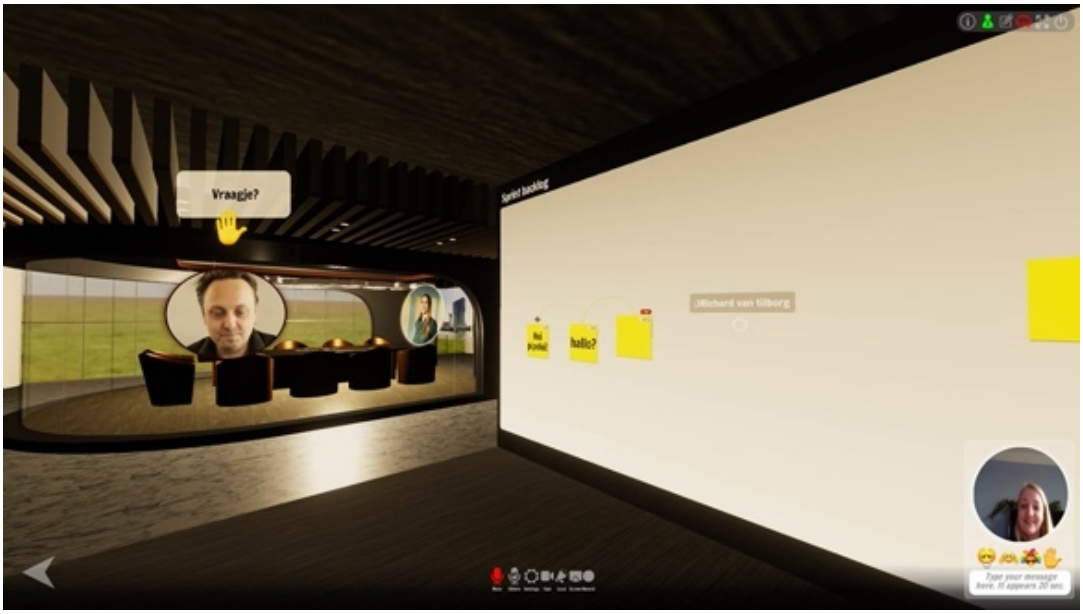

Figure 2: a screenshot of the adjacent room in the CoVince environment, with the post-it wall. *One of the users is raising a hand and using the text bubble function. Also, this user is pointing towards the wall with their right hand mouse button, so their name is visible on the wall with a small circle below.*

Four methods of measuring cognitive workload have been developed according to Zhou et al. (2018): *physiological*, *performance*, *behavioral*, and *subjective*. Physiological measurements were not used, as there are not in accordance with the global COVID-advice of social distancing. To measure performance with group task scores, the following group tasks were used during this experiment: a definitive solution task and a moral decision task. The requirements for both tasks were that they were group tasks that could be completed within the time span of around fifteen minutes with a scoring system. Therefore, the NASA moon survival task (hereinafter referred to as: the moon task) and the fallout shelter task (hereinafter referred to as: shelter task) were chosen. The moon task (Hall & Watson, 1970) is a task based on a situation in which the triad has to order items based on their usefulness to survive on the moon. This task is based on logical thinking and has a clear solution formulated by NASA experts. The score is based on a comparison of this ranking with the ranking of experts. The higher the score, the lower the performance. The shelter task (Simon et al., 2009) is also based on a hypothetical situation in which the triad has to order citizens in their usefulness to survive in a shelter during a war. This task doesn't have a definitive solution, because it is based on moral attitudes. Their score is based on the amount of unanimously ranked citizens within the time limit. The higher the score, the better the performance.

The behavioral measurement of cognitive workload was the use of plural first and third personal pronouns and the use of singular first and third personal pronouns (Sexton & Helmreich, 2000; Khawaja et al., 2012). This was assessed by recording the conversation of the participants. Afterwards, the conversation was transcribed and then analyzed with the Linguistic Inquiry and Word Count (LIWC). This software program analyzes a written text and calculated the pronouns as a percentage of total words used during the conversation. It was chosen instead of keeping a tally by hand to increase the trustworthiness of the results. It has been used in many linguistic studies (Zijlstra et al., 2004), and its validity has been demonstrated by Zijlstra et al. (2005), Bantum and Owen (2009), and Tausczik and Pennebaker (2010).

To measure subjective cognitive workload, the NASA Task Load Index (TLX) was used (NASA Task Load Index, 1986). This index is a multidimensional scale used to assess the subjective experience of cognitive workload after having executed a task or using a tool, such as software tools. It consists of 6 scales that are to be rated by the participants directly after completing the task. The score is calculated to be between 0 (low cognitive workload) and 100 (high cognitive workload). Its validity has been described by Hart and Staveland (1988) and numerous studies have employed it ever since, making it a standard in workload assessment (see Grier, 2005 for a review).

The results of the performance task and the NASA TLX were obtained with the survey website Qualtrics.

Design and Procedure

This experiment used a randomized factorial design (Kirk, 1982). Every software was matched with every task type and the order effect was accounted for by creating four conditions (see Figure 3). The triads were randomly assigned to one of these conditions, with two triads per condition.

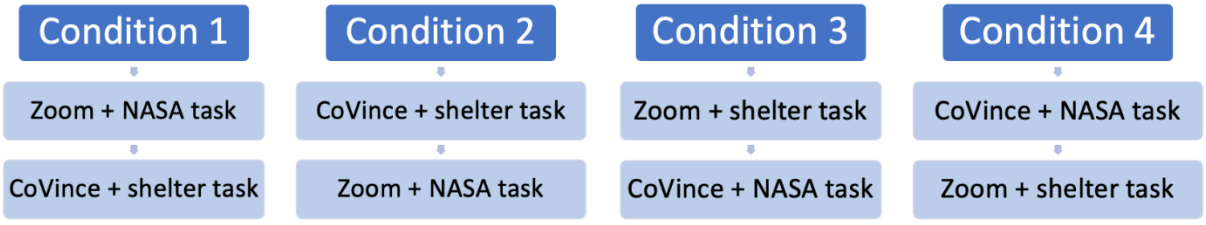

Figure 3: The factorial design scheme. *Every combination of software program and task as well as every order is included in the design. The triads are randomly assigned to these conditions.*

The participants signed up in triads consisting of three pre-acquainted colleagues. After application, they planned two time slots of 40 minutes with each other. Requirements were that these two time slots were not on the same day and were on the same moment of day (e.g. both in the morning). The protocol of each meeting in bullet points can be found in appendix A. The day before each of these time slots, they received an e-mail informing them about which software was to be used the next day and a link to the meeting in that software. They were instructed to open the link at the start of the time slot. In this meeting, the experimenter was present. To ensure similarity of provided information for all groups, The instructions given by the experimenter were scripted (Appendix B). A link was sent in the software which directed the participants to the online Qualtrics survey. This survey started with instructions and informed consent. Also, written instructions were given on the use of this particular software, while the participants were able to explore for five minutes. After this, the participants started the performance task that was assigned for that meeting (moon task or the shelter task). A timer was set for five minutes for them to individually contemplate the task and their own rankings. After this time, a fifteen minute time limit was set for them to discuss a ranking and reach group consensus. This time limit was based on previous research using the same task and a pilot study (discussed in the results) (Bottger & Yetton, 1987; Smolensky et al., 1990). During this period, their audio was recorded. One minute before the end of the individual time period and five minutes before the end of the group time period, a reminder was given about the remaining time. Within this fifteen minutes group time, all participants had to enter the ranking based on group consensus in the Qualtrics survey. For the shelter task, all participants of the triad had to enter the ranking decisions that were made with group consensus and to indicate when no unanimous decision was made about certain citizens in the list. When the time was up, the audio recording was stopped and all participants were kindly asked to fill in the NASA TLX individually in the Qualtrics survey. Age and average number of hours per week spend using videoconferencing tools were also assessed. In case this was the second meeting of that triad, they were asked which software was preferred in general and which was preferred to have these meetings, and an explanation was asked in a forced choice free text field. The total time taken from the start of the meeting until the end of the questionnaire was 40 minutes on average, so twice 40 minutes in total per triad.

Data analysis

To test for preconditions of employed tests (t-tests, ANOVAs, and correlations), respective preconditions were checked using Levene's test and Shapiro-Wilk test. If conditions were not met, non-parametric tests were run instead of their parametric counterparts. If any of the t-tests showed an insignificant result ($p > 0.05$), a Bayesian t-test was executed to evaluate the plausibility of the two compared datasets to be significantly similar.

Results

Pilot study

A pilot study was performed with one triad to ensure that the instructions and questionnaire were clear and that the time limits were appropriate for the tasks at hand. It showed that the experimental setup generally worked, but there were two problem areas: a ceiling effect in the shelter task and one of the NASA TLX scales. The shelter task score was 15 out of 15. As this limits the performance assessment, this was corrected by shortening the original time limit for both tasks from 7 minutes personal and 20 minutes group time to 5 minutes personal and 15 minutes group time. Furthermore, participants verbally stated that they were confused by the direction of one of the (translated) subscales of the NASA TLX, which is why this was inverted relative to the other questions of the NASA TLX. Afterwards, the scores of this scale were inverted to calculate the overall scores.

Loss of research data

During the experiment, 1 of 16 audio recordings failed. Due to software failure in the CoVince application, the moon task couldn't be completed for one triad, resulting in the loss of 1 datapoint. In total, 15 audio recordings, 8 shelter task results, 7 moon task results, and 48 NASA TLX scores were collected.

Task results

No significant difference was found between the moon task scores in the Zoom condition and in the CoVince condition ($t(3) = 2.47$, $p = 0.09$). With a Bayes Factor of $BF_{01} = 2.081$, there was anecdotal evidence for the hypothesis that the moon task scores of both software conditions differ from each other, of which the scores in the CoVince condition were (insignificantly) larger than the Zoom condition ($M = 41.33$, $SD = 7.02$ and $M = 30$, $SD = 4.32$, respectively). So, with a higher score meaning a worse performance, the participants scored descriptively better on the moon task in the Zoom condition than in the CoVince condition.

A ceiling effect was found for the performance scores on the shelter task: all groups scored 15 out of 15, except for one group scoring 9 out of 15. Because of the lack of variation within the two groups, a statistical analysis to test the randomness of the variation was futile.

Linguistic results

Both the CoVince and the Zoom condition showed a significant difference in the use of these singular and plural personal pronouns; however, the percentage of singular first and third personal pronouns was bigger than the percentage of plural first and third personal pronouns in both cases. In the CoVince condition, the results were $t(13) = 3.38$, $p = 0.005$ with a singular percentage of $M = 4.80\%$ ($SD = 0.94\%$) and a plural percentage of $M = 3.25\%$ ($SD = 0.90\%$); in the Zoom condition, the results were $t(10) = 3.83$, $p = 0.003$ with a singular percentage of $M = 5.18\%$ ($SD = 0.92\%$) and a plural percentage of $M = 3.07\%$ ($SD = 1.34\%$). So, the participants used significantly more singular first and third personal pronouns than plural in both software conditions.

Like the software conditions, the shelter and moon task showed a significant higher percentage of singular first and third personal pronouns over plural first and third personal pronouns. For the shelter task, the results were $t(11) = 3.28$, $p = 0.007$ with a singular percentage of $M = 4.57\%$ ($SD = 0.60\%$) and a plural percentage of $M = 3.26\%$ ($SD = 0.95\%$). For the moon task, the results were $t(9) = 4.30$, $p = 0.002$ with a singular percentage of $M = 5.18\%$ ($SD = 0.93\%$) and a plural percentage of $M = 2.52\%$ ($SD = 1.20\%$). Therefore, the participants used significantly more singular first and third personal pronouns than plural in both task conditions.

NASA TLX results

Cronbach's alpha was 0.82, which is an indication of a robust internal consistency. If the confusing question with the inversed scale would be removed, the alpha value would go up to 0.85. As this increase is relatively small, the question was included in the analysis.

NASA TLX ratings did not differ significantly between the software conditions based on a two-way ANOVA ($F(1, 44) = 0.09$, $p = 0.76$). On the other hand, a significant effect was found of the tasks executed on the NASA TLX score: the scores were significantly higher after executing the shelter task than after executing the moon task ($F(1, 44) = 11.95$, $p < 0.001$) (Figure 4). The average score after completing the shelter task was 44.62 ($SD = 12.31$), and 30.69 ($SD = 16.33$) for the moon task. Hence, participants rated their own cognitive workload higher on the NASA TLX scales after having completed the shelter task than after having completed the moon task.

There also was a significant interaction effect between the effect of the software application used and the task executed on the NASA TLX scores ($F(1,44) = 5.24$, $p < 0.01$) (Figure 4). Tukey's HSD post hoc tests were executed and showed that there is a significant difference in NASA TLX scores between the CoVince condition during the shelter task and the CoVince condition during the moon task ($p = 0.001$) (Figure 4).

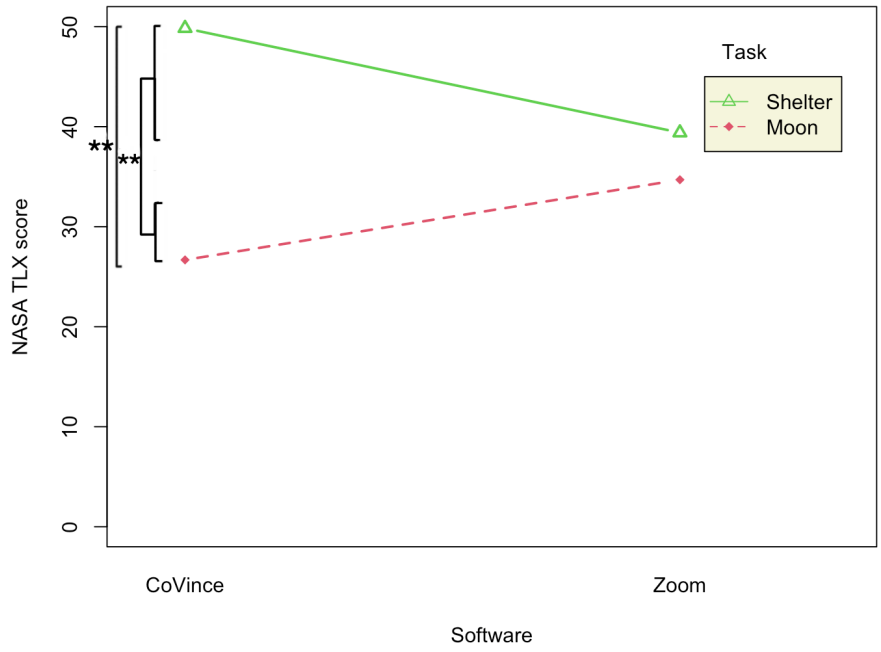

Figure 4: The interaction effect between the software and the task on the TLX score. *The software applications are shown on the x-axis, the TLX score is shown on the y-axis and both tasks are shown with a line. ** , p < 0.001*

Preferences

To the question 'Which software did you prefer to hold a meeting with?' at the end of every second meeting, 14 participants answered Zoom, 8 answered CoVince, and 5 participants didn't have a preference. The question 'Which software did you prefer to use in general?' 18 participants answered Zoom, 7 participants answered CoVince, and 2 participants didn't have a preference. The main argumentation in favor of Zoom was that it's easier to use, it's more comprehensive, and it's easier to see the other participants. No argumentation against Zoom was given by any participant. The main argumentation in favor of the CoVince software, was that the post-it board is a great addition, the spatiality benefits conversation, and it is 'more fun and inspiring to use'. Argumentation against the CoVince was mostly that there were too many bugs which hinders the use of the program, the software was too

weighty for laptops which made the laptops heat up and slow down, and executing the tasks verbally without using the software was quicker. Generally, some participants noted that the use of the software is dependent on the task one wants to execute: the CoVince software was deemed to be more useful for 'fun' and light tasks, Zoom was preferred for a quick and goal-oriented task.

Explorative analyses

Because the data of this research enable a more explorative data analysis, several other analyses were executed to explore other effects in the data that was collected. These analyses, due to the lack of references backing up these conclusions, give rise to larger-scale studies to give more robust analyses and conclusions.

Plural first and third person pronoun use

As an addition to the linguistic analyses done in this research, it was also investigated whether an interaction effect of task and software application would be apparent in the linguistic data. To investigate this, the proportion of plural first and third personal pronouns was taken of all first and third pronouns (singular and plural) for each condition. A two-way ANOVA showed a significant interaction effect of software and task on the proportion of plural first and third pronouns ($F(1, 11) = 9.57$, $p = 0.01$). Tukey's HSD post hoc tests were executed and showed two significant differences: between the Zoom – moon and the CoVince – moon condition ($p = 0.03$) and between the Zoom – shelter and the Zoom moon condition ($p = 0.02$). In Figure 5, these results are depicted.

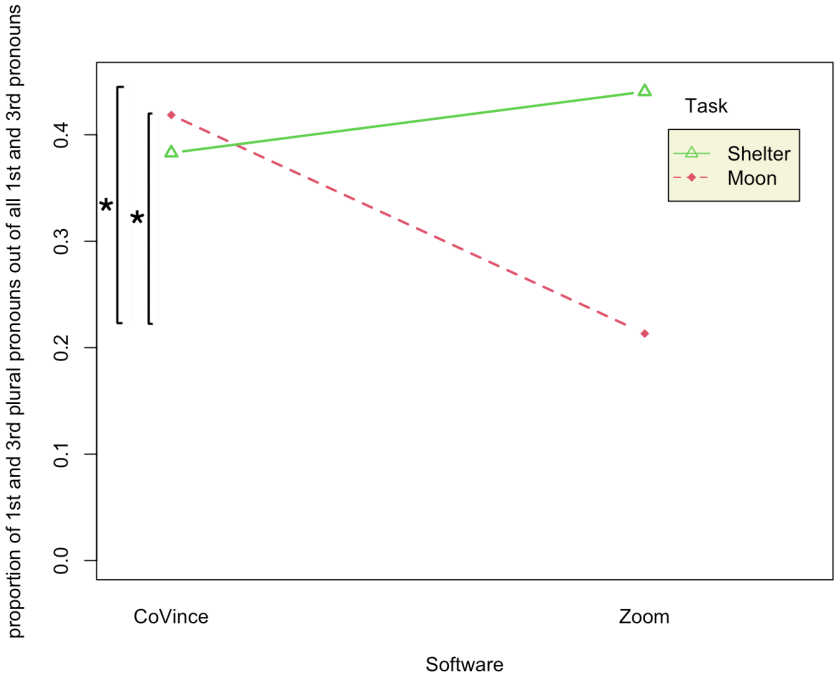

Figure 5: The interaction effect between the software and the task on the linguistic pronouns proportions*. The software applications are shown on the x-axis, the proportion plural first and third pronouns of both singular and plural first and third pronouns is shown on the y-axis and both tasks are shown with a line. * , p < 0.01*

Correlations

To investigate whether differences in age and experience influence the NASA TLX results, the correlation between weekly hours using Zoom and age, NASA TLX score and age, and weekly hours using Zoom and age were investigated. No significant correlations were found. These findings may be of relevance for future research into cognitive workload imposed by online conferencing tools. The results are depicted in Appendix C.

Discussion

The aim of this thesis project was twofold. The first aim was to develop an experimental design to benchmark a videoconferencing application (here: Zoom) against a Virtual Reality application (here: CoVince) in terms of cognitive workload which could be executed under COVID-restrictions. Participants held meetings in triads with both applications. Several measurements of cognitive workload were conducted, based on previous research: *performance scores* on two different tasks, *linguistic measurements,* and *subjective* scores on the NASA TLX. The here developed paradigm and tested tasks might serve as a blueprint for future studies into which online communication software programs are best used for working from home with as little imposed cognitive workload as possible. The second goal was to provide first insights into whether one of these two software programs outperforms the other in terms of imposed cognitive workload. The results show that the participants didn't score significantly different on the moon task in both software conditions. The shelter task showed a ceiling effect. The linguistic measurements showed significant effects, but inversed from the findings of previous research (Sexton & Helmreich, 2000; Khawaja et al., 2012), i.e. singular first and third personal pronouns were used significantly more than plural first and third personal pronouns in all conditions, independent of the NASA TLX scores. NASA TLX scores didn't differ significantly between the CoVince and Zoom conditions. So, these measurements give no clear indication of one application outperforming the other in terms of cognitive load. After an explorative linguistic analysis, it was found that the participants proportionally used more plural first and third personal pronouns in the CoVince condition than in the Zoom condition for the moon task. Also, NASA TLX scores were significantly lower for the moon task than the shelter task in the CoVince condition. These results don't match our prediction that the increased use of first and third plural pronouns is correlated with increased NASA TLX scores, based on the research of Sexton and Helmreich (2000) and Khawaja et al. (2012). It is speculative, but it is plausible to assume that this increased use of first and third personal pronouns and decreased NASA TLX scores could be interpreted as being indicative of improved teamwork in the CoVince condition when the moon task is executed compared to the other conditions. Using 'we' and 'ours' more than 'I' and 'mine' in combination with decreased subjective workload seem to be plausible indications of the feeling of group cohesion and group responsibility.

When asked about preferences, Zoom was preferred by most for having meetings and in general. Arguments were that it was easier to use and more comprehensive. Arguments in favor of CoVince were that the spatiality helps and it was more inspiring to use. It is possible that this preference is not based on imposed cognitive workload by feasibility features, but on the acceptability. Zoom is simply used more frequently and by more participants, and therefore more well-known. Based on all results, it could be said that both tools have their strengths. Zoom is deemed to be more useful for goal-oriented and meetings with a focus on moral decision making. CoVince, on the other hand, may be more suited for brainstorming as a team and meetings focused on creativity.

These results are partly in line with the hypothesis based on previous studies. These studies demonstrated that a lack of spatiality in videoconferencing communication limit nonverbal communication strategies (Gaver, 1992; Sellen, 1992; Hauber et al., 2006; Sirkin, 2011). The addition of spatiality has been demonstrated to decrease cognitive workload (Monk & Gale, 2002). These findings were matched with the findings in this experiment, but only when the moon task was executed. It could be that the combination of the definitive solution nature of this task and the CoVince application causes this effect; participants trusted each other's knowledge on the matter instead of discussing opinions like in the shelter task, and the CoVince application might inspire creativity in using this knowledge. As most office meetings rely at least partly on each individual bringing their own knowledge and expertise to the table, these meetings might benefit from using CoVince instead of Zoom. It should be mentioned that the difference in spatiality between the CoVince and Zoom application is probably not the only feasibility aspect that causes this difference in the results: the CoVince application also has an elaborate design of both the room and the environment, a post-it board, and multiple emoticon and chat functions.

Another explanation of these findings is that the measurements of cognitive workload are not sufficient enough to reflect imposed cognitive load validly. Other possible causes of increased cognitive workload in Zoom not related to spatiality are namely also present in CoVince, such as lag time in verbal responses (Roberts & Francis, 2013), the continuous awareness of being watched (Jiang, 2020), and possibly increasing heuristic processing (Chaiken & Ledgerwood, 2011). On the other hand, the enlarged display of the

heads of others causing fatigue (Reeves et al., 1999), is more prominent in Zoom than in CoVince. An evaluation of the measures used in this research is given below.

Evaluation of the research design

To allow further investigations building upon the here developed experimental design and setup, the measurements, software programs and participants of this experiment are discussed in the following section.

Measurements

Cognitive workload, defined as "*a multidimensional construct that represents the workload that performing a particular task imposes on the cognitive system of a learner*" by Paas and van Merriërboer (1994) is per definition a broad term. Therefore, it is hard to measure this concept. This might reflect weak measurement tools or a definition that is too vague. Because of the vagueness of this concept, the measurements in this research seem to be indications of more than just cognitive workload, for example communication, collaboration, and personal skills. Maybe the addition of physiological measures could contribute to a more defined measurement. Here, each measurement used in this experiment will be evaluated on its reflection of cognitive workload.

The shelter and moon task served as a performance measure of cognitive workload. The shelter task had an obstructing limitation: the ceiling effect. It seemed like the participants were more keen on finishing the whole ranking within the time, than ensuring that the ranking was compatible with their own opinions. This was the case in both the pilot study and the experiment itself, while both had different time limits. Also, this task significantly increased the subjective cognitive workload compared to the moon task condition shown in the NASA TLX; therefore, when focusing on the cognitive workload differences in the software, it might be better for equal variables to choose one task, or find a moral decision task with an imposed cognitive workload more comparable to the moon task. A general point of criticism might be that these performance measures of cognitive workload are not an indication of cognitive workload alone. Apart from being an indication of the workload that performing this task imposes on the cognitive system of the learner, performance scores on these tasks are also an indication of, for example, the reasoning orientation of the group (exchanging facts instead of sticking to their positions) (Innami,

1994), team conscientiousness, and agreeableness (Kimura & Kottke, 2009). Therefore, having only these performance scores to draw conclusions on the cognitive workload of the participants may be a simplification of all variables in place. On the other hand, the value of these performance scores should not be overlooked: the other variables that influence these performances scores might be valuable indications of the acceptability and effectiveness of videoconferencing applications.

The same can be said about the linguistic measurements of the singular first and third personal pronouns and the plural first and third personal pronouns. The use of pronouns is probably a presentation of much more than cognitive workload, such as teamwork and task type. For instance, it has been demonstrated that the use of personal pronouns are a reflection of social hierarchies (Kacewicz et al., 2014) and negative team interactions (Yilmaz, 2014). Additionally, the tasks in this experiment entailed an individual ranking before discussing a group ranking, therewith increasing the use of singular pronouns. This may differ for other task types without an individual evaluation. Moreover, the studies of Sexton and Helmreich (2000) and Khawaja et al. (2012) on which these measurements were based, were conducted in English. This research was conducted in Dutch. Even though English and Dutch are both West Germanic languages and therefore related (Harbert, 2006), it may be possible that the results of these studies are generalizable to Dutch. Further research into this generalization needs to be done before these findings can be trustworthily used in Dutch studies concerning cognitive workload.

Lastly, the NASA TLX was used to measure subjective cognitive workload. After having inversed the last scale, it seemed like the participants understood the scales. This was shown with the Cronbach's alpha. It is a deviation of the standard NASA TLX; however, this may be attributed to the effect of translation. This measure of cognitive workload as reviewed by Grier (2015) showed its functionality in this experiment as well.

Software programs

The software programs used in this research were Zoom and CoVince. The Zoom software was familiar with all the participants and presented no problems concerning the use of the software: no bugs or software problems arose during the experiment.

Because the CoVince software is relatively new, a big influence on the results of this experiment could have been the excitement that comes with using new applications (Ahn & Shin, 2015). This excitement of using new and unknown features may increase the requisite cognitive capacity of handling the software application, therewith increasing cognitive workload (Lin et al., 2019). Using new software also creates a big learning curve. Zoom was known to all participants, so participants quickly used features and navigated easily. This was not the case for the CoVince platform; even after having 5 minutes to gain familiarity the navigation and features of the software, this didn't guarantee dexterity of using these features during the experiment. So possibly, when this software is more known to its users and no extra windows are necessary for executing a task (e.g. brainstorming or scrumming), this software may be more beneficial. An aspect that adds to that difficulty in navigation and features of the software, is that the version used in this experiment was still in an early phase. It was prone to a great deal of bugs (e.g. screen freezes, audio issues, navigational problems). These bugs were solved each time they arose, therewith using the feedback of the participants to improve the software. Hence, it could be assumed that the general user experience improved throughout the experiment.

Participants

Having no age or experience restrictions when looking for colleague participants was an important aspect of participant recruitment, as all age groups of the working class are using videoconferencing tools when meeting from home and it is therefore informative to include all experiences. As the explorative results show, age and experience of using a software application don't correlate with the NASA TLX results. This means that the results of this questionnaire are not skewed by age or experience. This relationship between age and the NASA TLX results was also demonstrated before (Mouzé-Amady et al., 2013). This experiment can therefore be reliably used for heterogenous participant databases. Further, more extensive research is needed to back up these claims.

Using triads of participants appears suitable for these tasks and software applications, as they are big enough for multiple opinions to be expressed, while also being orderly. For future research, it is necessary to increase the number of groups when conducting comparative research of software applications into cognitive workload in order to increase the power of the research.

Future research

To begin with, future research might benefit from a more specific definition of the term cognitive workload or measurements that reflect cognitive workload more specifically. The addition of physiological measurements might help to reflect cognitive workload more clearly.

Furthermore, is advised that future research into cognitive workload differences in online conferencing applications consider the following adjustments. Firstly, as the shelter task demonstrated a clear ceiling effect, this task is not suitable as a performance measure for cognitive workload. Secondly, having multimodal measurements of cognitive workload may increase the validity of the conclusions, but this only applies if all measurements are robust indications of cognitive workload, specifically. Therefore, the linguistic measurements used in this research may not be useful, unless more research is done into the cause(s) of this linguistic effect shown by Sexton and Helmreich (2000) and Khawaja et al. (2012). For example, is it only cognitive workload, or also other causes like team work and group hierarchy that cause this linguistic effect? And what is the extent to which this measurement is influenced by task type? Is this linguistic effect also seen in languages other than English?

Research into the effect of feasibility aspects differing between Zoom and CoVince other than spatiality will also create more opportunity to improve the design of videoconferencing tools. Features like a seeing all video figures in a more natural environment, the presence of a post-it board and opting whether to see your own face or not are all feasibility features differing between the applications of which it could be informative to see the impact on the cognitive workload of the user. Other than feasibility, the other aspects of a good app are also possible focuses of future research: the acceptability, effectiveness, and costs of online conferencing applications. Because videoconferencing tools like Zoom are widely used, the acceptability is extremely high. So even if the benchmark application shows flaws, it is hard to break the paradigm of that application. Research into how this paradigm might be broken and how acceptability of new applications could be increased, could be interesting for surging applications such as CoVince. The effectiveness of new applications are influenced by the feasibility features of the application and the purpose of the user. So, which purposes can be defined? And which

feasibility features match these purposes the best to increase effectiveness? Lastly, it is informative to include the costs of online conferencing applications in all comparative studies, as this is paramount in the decision making of an employer on which online conferencing software application to purchase.

Conclusion

I here provide a research design to benchmark different conferencing software applications with each other concerning cognitive workload on users. The here developed paradigm and tested tasks might serve as a blueprint for future studies into which online communication software programs are best used for working from home with as little imposed cognitive workload as possible. First results into the differences in imposed cognitive workload of the videoconferencing application Zoom relative to the Virtual Reality application CoVince have shown that both applications have their own strengths. CoVince outperforms Zoom on the NASA TLX scores, but only when the moon task is executed. Explorative linguistic results show that plural first and third plural pronouns are used proportionally more in CoVince than Zoom when the moon task is executed. These two findings may suggest that cognitive workload is less and teamwork is enhanced in CoVince when the Zoom task is executed. Based on this conclusion in combination with the written feedback of the participants, it could be said that both tools have their strengths; Zoom was deemed to be more useful for goal-oriented and meetings with a focus on moral decision making. CoVince, on the other hand, might be more suited for brainstorming as a team and meetings focused on creativity.

References

Adams, S. J., Roch, S. G., & Ayman, R. (2005). Communication medium and member familiarity: The effects on decision time, accuracy, and satisfaction. *Small group research*, *36*(3), 321-353.

Ahn, D., & Shin, D. H. (2015). Differential effect of excitement versus contentment, and excitement versus relaxation: Examining the influence of positive affects on adoption of new technology with a Korean sample. *Computers in Human Behavior*, *50*, 283-290.

Allen, M., Sargeant, J., Mann, K., Fleming, M., & Premi, J. (2003). Videoconferencing for practice-based small-group continuing medical education: Feasibility, acceptability, effectiveness, and cost. *Journal of Continuing Education in the Health Professions*, *23*(1), 38-47.

Aries, M. B., Veitch, J. A., & Newsham, G. R. (2010). Windows, view, and office characteristics predict physical and psychological discomfort. *Journal of environmental psychology*, *30*(4), 533-541.

Bantum, E. O. C., & Owen, J. E. (2009). Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological assessment*, *21*(1), 79.

Bick, A., Blandin, A., & Mertens, K. (2020). Work from home after the COVID-19 Outbreak.

Bottger, P. C., & Yetton, P. W. (1987). Improving group performance by training in individual problem solving. *Journal of Applied Psychology*, *72*(4), 651.

Businesswire (2020). COVID-19 Outbreak: Videoconferencing Demand Rises due to Social-Distancing – Researchandmarkets.com. Available at: https://www.businesswire.com/news/home/20200507005631/en/COVID-19-

Outbreak-Video-Conferencing-Demand-Rises-due-to-Social-Distancing---ResearchAndMarkets.com, accessed February 26th, 2021.

Chaiken, S., & Ledgerwood, A. (2011). A theory of heuristic and systematic information processing. *Handbook of theories of social psychology: Volume one*, 246-166.

Corporate vision, n.d. Technology Innovator Awards. Available at: https://www.cv-magazine.com/winners/2019-covince-innovations-b-v/, accessed March 10th, 2021.

CoVince (2021). Download CoVince. Available at: https://covince.com/download, accessed February 26th, 2021.

Ferran, C., & Storck, J. (1997). Truth or Deception: The Impact of Videoconferencing for Job Interviews. In *Proceedings of the Eighteenth International Conference on Information Systems (ICIS 97), Atlanta, Georgia*.

Ferran, C., & Watts, S. (2008). Videoconferencing in the field: A heuristic processing model. *Management science*, *54*(9), 1565-1578.

Fosslien, L., & Duffy, M. W. (2020). How to combat zoom fatigue. *Harvard Business Review*, *29*.

Gaver, W. W. (1992). The affordances of media spaces for collaboration. In *Proceedings of the 1992 ACM conference on Computer-supported cooperative work* (pp. 17-24).

Grier, R. A. (2015, September). How high is high? A meta-analysis of NASA-TLX global workload scores. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*(Vol. 59, No. 1, pp. 1727-1731). Sage CA: Los Angeles, CA: SAGE Publications.

Hall, J., & Watson, W. H. (1970). The effects of a normative intervention on group decision-making performance. *Human relations*, *23*(4), 299-317.

Harbert, W. (2006). *The Germanic Languages*. Cambridge University Press.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.

Hauber, J., Regenbrecht, H., Billinghurst, M., & Cockburn, A. (2006). Spatiality in videoconferencing: trade-offs between efficiency and social presence. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work* (pp. 413-422).

Holshue, M. L., De Bolt, C., Lindquist, S., Lofy, K. H., Wiesman, J., Bruce, H., ... & Pillai, S. K. (2020). First case of 2019 novel coronavirus in the United States. *New England Journal of Medicine*.

Holzapfel, B. (2020). New Teams features add creative ways to engage students . Available at: https://educationblog.microsoft.com/en-us/2020/07/new-teams-features-add-creative-ways-to-engage-students/, accessed March 24th, 2020.

Innami, I. (1994). The quality of group decisions, group verbal behavior, and intervention. *Organizational Behavior and Human Decision Processes*, *60*(3), 409-430.

Jiang, M. (2020). The reason Zoom calls drain your energy. *BBC, April*, *22*, 179.

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2014). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, *33*(2), 125-143.

Kendon, A. Some functions of gaze-direction in social interaction. *Acta Psychologica*, *32*(1967), 1-25.

Khawaja, M. A., Chen, F., & Marcus, N. (2012). Analysis of collaborative communication for linguistic cues of cognitive load. *Human factors*, *54*(4), 518-529.

Kimura, S., & Kottke, J. L. (2009). Cognitive ability and personality can predict team productivity but not team synergy. In *Annual Conference of the Association for Psychological Science, San Francisco, California*.

Kirk, R. E. (2012). Experimental design. *Handbook of Psychology, Second Edition*, *2*.

Lee, J. (2020). A neuropsychological exploration of zoom fatigue. *Psychiatric Times. https://www. psychiatrictimes. com/view/psychological-exploration-zoom-fatigue*.

Lewis, D., Tranter, G., & Axford, A. T. (2009). Use of videoconferencing in Wales to reduce carbon dioxide emissions, travel costs and time. *Journal of Telemedicine and Telecare*, *15*(3), 137-138.

Lin, H. C. S., Yu, S. J., Sun, J. C. Y., & Jong, M. S. Y. (2019). Engaging university students in a library guide through wearable spherical video-based virtual reality: effects on situational interest and cognitive load. *Interactive Learning Environments*, 1-16.

Littlepage, G., Robison, W., & Reddington, K. (1997). Effects of task experience and group experience on group performance, member ability, and recognition of expertise. *Organizational behavior and human decision processes*, *69*(2), 133-147.

Miller, P. A., Huijbregts, M., French, E., Taylor, D., Reinikka, K., Berezny, L., ... & Harvey, M. (2008). Videoconferencing a stroke assessment training workshop: effectiveness, acceptability, and cost. *Journal of Continuing Education in the Health Professions*, *28*(4), 256-269.

Monk, A. F., & Gale, C. (2002). A look is worth a thousand words: Full gaze awareness in video-mediated conversation. *Discourse Processes*, *33*(3), 257-278.

Morawska, L., & Cao, J. (2020). Airborne transmission of SARS-CoV-2: The world should face the reality. *Environment international*, *139*, 105730.

Moulton-Perkins, A., Moulton, D., Cavanagh, K., Jozavi, A., & Strauss, C. (2020). Systematic review of mindfulness-based cognitive therapy and mindfulness-based stress reduction via group videoconferencing: Feasibility, acceptability, safety, and efficacy. *Journal of Psychotherapy Integration*.

Mouzé-Amady, M., Raufaste, E., Prade, H., & Meyer, J. P. (2013). Fuzzy-TLX: using fuzzy integrals for evaluating human mental workload with NASA-Task Load indeX in laboratory and field studies. *Ergonomics*, *56*(5), 752-763.

Miner Jr, F. C. (1984). Group versus individual decision making: An investigation of performance measures, decision strategies, and process losses/gains. *Organizational Behavior and Human Performance*, 33(1), 112-124.

NASA Task Load Index (1986). Paper-and-Pencil version. *Moffett Field, CA: NASA-Ames Research Center, Aerospace Human Factors Research Division*.

Nguyen, D. T., & Canny, J. (2007). Multiview: improving trust in group video conferencing through spatial faithfulness. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 1465-1474).

Okta (2021). Businesses at Work. Available at: https://www.okta.com/sites/default/files/2021-02/Businesses-at-Work-2521.pdf, accessed March 10th, 2021.

Ong, D., Moors, T., & Sivaraman, V. (2014). Comparison of the energy, carbon and time costs of videoconferencing and in-person meetings. *Computer communications*, *50*, 86-94.

Oseland, N. (2009). The impact of psychological needs on office design. *Journal of Corporate Real Estate*.

Paas, F. G., & Van Merriënboer, J. J. (1994). Instructional control of cognitive load in the training of complex cognitive tasks. *Educational psychology review*, *6*(4), 351-371.

Reeves, B., Lang, A., Kim, E. Y., & Tatar, D. (1999). The effects of screen size and message content on attention and arousal. *Media psychology*, *1*(1), 49-67.

Roberts, F., & Francis, A. L. (2013). Identifying a temporal threshold of tolerance for silent gaps after requests. *The Journal of the Acoustical Society of America*, *133*(6), EL471-EL477.

Sellen, A. J. (1992). Speech patterns in video-mediated conversations. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 49-59).

Sexton, J. B., & Helmreich, R. L. (2000). Analyzing cockpit communications: the links between language, performance, error, and workload. *Human Performance in Extreme Environments*, 5(1), 63-68.

Simon, S. B., Howe, L. W., & Kirschenbaum, H. (2009). *Values clarification*. Grand Central Publishing.

Sirkin, D., Venolia, G., Tang, J., Robertson, G., Kim, T., Inkpen, K., ... & Sinclair, M. (2011, September). Motion and attention in a kinetic videoconferencing proxy. In *IFIP Conference on Human-Computer Interaction* (pp. 162-180). Springer, Berlin, Heidelberg.

Smith, H. J., & Neff, M. (2018). Communication behavior in embodied virtual reality. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1-12).

Smolensky, M. W., Carmody, M. A., & Halcomb, C. G. (1990). The influence of task type, group structure and extraversion on uninhibited speech in computer-mediated communication. *Computers in Human Behavior*, 6(3), 261-272.

Stone, N. J., & English, A. J. (1998). Task type, posters, and workspace color on mood, satisfaction, and performance. *Journal of Environmental Psychology*, *18*(2), 175-185.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, *29*(1), 24-54.

Trueman C. (2020). Pandemic leads to surge in video conferencing app downloads. Available at: https://www.computerworld.com/article/3535800/pandemic-leads-to-surge-in-video-conferencing-app-downloads.html, accessed February 26th, 2021.

Vertegaal, R. (1997). Conversational awareness in multiparty VMC. In *CHI'97 Extended Abstracts on Human Factors in Computing Systems* (pp. 6-7).

Vertegaal, R. (1999). The GAZE groupware system: mediating joint attention in multiparty communication and collaboration. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* (pp. 294-301).

Wiederhold, B. K. (2020). Connecting through technology during the coronavirus disease 2019 pandemic: Avoiding “Zoom Fatigue”.

Wolf, C. R. (2020). Virtual platforms are helpful tools but can add to our stress. *Psychology Today*.

World Health Organization, (2021). Weekly epidemiological update – 25 Febuary 2021. Available at: https://www.who.int/publications/m/item/covid-19-weekly-epidemiological-update. Accessed February 26th, 2021.

Xiao, Y. M., Wang, Z. M., Wang, M. Z., & Lan, Y. J. (2005). The appraisal of reliability and validity of subjective workload assessment technique and NASA-task Load index. *Zhonghua lao dong wei sheng zhi ye bing za zhi= Zhonghua laodong weisheng zhiyebing zazhi= Chinese journal of industrial hygiene and occupational diseases*, *23*(3), 178-181.

Yilmaz, G., & Peña, J. (2014). The influence of social categories and interpersonal behaviors on future intentions and attitudes to form subgroups in virtual teams. *Communication Research*, *41*(3), 333-352.

Zhou, J., Yu, K., Chen, F., Wang, Y., & Arshad, S. Z. (2018). Multimodal behavioral and physiological signals as indicators of cognitive workload. In The Handbook of Multimodal-Multisensor Interfaces: *Signal Processing, Architectures, and Detection of Emotion and Cognition*-Volume 2 (pp. 287-329).

Zijlstra, H., Van Meerveld, T., Van Middendorp, H., Pennebaker, J. W., & Geenen, R. (2004). De Nederlandse versie van de 'linguistic inquiry and word count'(LIWC). *Gedrag Gezond*, *32*(4), 271-281.

Zijlstra, H., Van Middendorp, H., Van Meerveld, T., & Geenen, R. (2005). Validiteit van de Nederlandse versie van de Linguistic Inquiry and Word Count (LIWC). *Nederlands Tijdschrift voor de Psychologie en haar Grensgebieden*, *60*(3), 50-58.

Zoom (2021). Zoom. Available at: https://zoom.us. Accessed February 26th, 2021.

Appendix A: protocol experiment

The protocol from finding participants to setting up the meetings.

1. Reach out to possible participants and send them the information letters.
2. Have them plan two moments to have the meetings. A day before both meetings, send them a reminder with the date, time, and the link to the right software meeting.
3. On the day of the meeting, preparations need to be made. Make sure that the questionnaires and assignments are ready, the respective software is running smoothly, the audio recording device is working, and a stopwatch is present.
4. Be present in the meeting 5 minutes before the planned moment, to receive early participants. When all are present, follow the script (Appendix B).
5. During the experiment, send the participants the link to the questionnaire (this was done with Qualtrics). In Zoom, this is done by sending the URL in the chat function. In CoVince, this is done by having a post-it on the wall with the URL. Within Qualtrics, they give their informed consent, read the instructions on navigation about the software they're using, and read the instructions on the task they're executing that meeting. They are given 5 minutes to practice the navigation in both software applications.
6. They start the task after this navigation practice. Make sure that the timer is running and the audio is recording when they are executing the task. Don't forget to give a 5-minutes-left reminder at the end of the task time and 1-minute left reminder at the end of the personal time. Also, stress that all participants have to enter the group answers to the task before the timer runs out.
7. It is important as a researcher to be in the background; don't give away answers and don't participate in the group discussions. Put yourself on mute and only unmute when a questions is asked directly. Only questions about the software and the execution of the task can be answered; no answers on the content of the tasks can be answered.
8. After the task is done, the participants will fill in the NASA TLX on their own in Qualtrics.
9. When this is their second meeting, they will also fill in their personal preferences for either one of the software programs and their argumentation.

10. Thank the participants for their participation. In the case that they executed the moon task, send them the PDF with the answers afterwards if they are interested in them.
11. For data analysis, first transcribe the audio by hand. Then, run the text through the LIWC (use a dictionary with the right language: in this case, Dutch). Calculate the NASA TLX score per participant and the task score per group.

Appendix B: script

Hallo, mijn naam is Cato Zantman en ik ben een masterstudente aan de studie Applied Cognitive Psychology. Deze studie focust zich op de interactie tussen de mens en haar omgeving en hoe we deze omgeving zo geschikt mogelijk kunnen maken gebaseerd op het functioneren van de mens.

Op het moment ben ik mijn scriptie-onderzoek aan het uitvoeren. Ik probeer verschillende manieren van online vergaderen te onderzoeken en daarbij te kijken naar welke manier het prettigst is om te gebruiken. Zoals jullie wel hebben gemerkt, zitten we in een pandemie waarbij we opeens enorm veel gebruik maken van online vergaderen. Jullie hebben misschien ook wel gemerkt dat dat erg vermoeiend kan zijn, soms wel vermoeiender dan face to face vergaderen. Om die reden ben ik dit onderzoek gestart en dit is waar jullie in het spel komen. Ik ga met jullie en nog een aantal andere groepjes twee meetings houden: eentje via de software Zoom en eentje via de software CoVince. In beide meetings gaan jullie een groepsopdracht doen die op elkaar lijken maar niet precies hetzelfde zijn, en dan ga ik aan de hand van een aantal metingen kijken in welke situatie dit nou het prettigst verliep. Is dit voor iedereen duidelijk?

De rode draad door deze meeting heen zal een online vragenlijst zijn.

- Zoom: ik zal zo meteen in de chat van deze meeting de link sturen naar die online vragenlijst.
- CoVince: jullie hebben daarnet in een korte tutorial te zien gekregen hoe je rondloopt in deze ruimte. Nu mogen jullie met mij meelopen naar de zaal hiernaast, bij het post-it bord. Op dit bord hangt een post-it genaamd ‘link vragenlijst’. Dit zal jullie straks leiden naar de desbetreffende vragenlijst.

In deze vragenlijst zien jullie eerst een paar informatiepagina’s en daarna het informed consent. Nadat jullie dit rustig door hebben kunnen lezen, zal er gevraagd worden welke taak jullie gaan doen. Zodra jullie daar zijn aangekomen, mag je even je hand opsteken. Op het moment dat jullie alle drie gereed zijn, zal de taak beginnen. Deze is als volgt vormgegeven: er wordt een situatie omschreven waarin jullie je als team bevinden. Vervolgens zie je een lijst van 15 objecten of personen die in deze situatie voorkomen, aan jullie de taak om een unanieme ranking te vormen van deze objecten of personen. Eerst hebben jullie 5 minuten om voor jezelf te bedenken welke ranking je zou willen, daarna krijgen jullie 15 minuten de tijd om de gezamenlijke ranking te bedenken. Het is de

bedoeling dat je uiteindelijk de gezamenlijke ranking in de online vragenlijst invult. 1 minuut voor het einde van de persoonlijke tijd en 5 minuten voor het einde van de gezamenlijke tijd zal ik een reminder geven.

Na de groepstaak krijgen jullie een korte vragenlijst over je individuele ervaringen die je in mag vullen. Zelf sta ik de rest van de meeting op mute met m'n video uit zodat ik geen afleiding ben, maar ik ben wel continu aanwezig; mochten jullie vragen hebben, dan kan je me roepen en dan beantwoord ik je vraag. Ik kan alleen vragen beantwoorden over de uitvoering van de opdracht, niet over de inhoud.

- Shelter taak: deze taak is er niet op gebouwd dat de volledige ranking van 15 personen binnen de tijd unaniem afgerond wordt. Het is dus mogelijk dat jullie een aantal mensen een plek hebben kunnen geven, maar het niet eens worden over de rest. Deze personen kan je in de online vragenlijst aangeven als 'geen unanieme beslissing'.

Is dit voor iedereen duidelijk?

Dan is hier de link naar de vragenlijst en dan wens ik jullie veel succes.

Appendix C: explorative analyses

To test whether age influenced the weekly amount of hours that software programs are used, a correlation test was executed between age and the weekly hours spent on Zoom (no participants used the CoVince software regularly). Because the assumption of normality was not met, the non-parametric Spearman correlation test was executed. No significant correlation was found (S = 1684, p = 0.21). The data is shown in Figure 6.

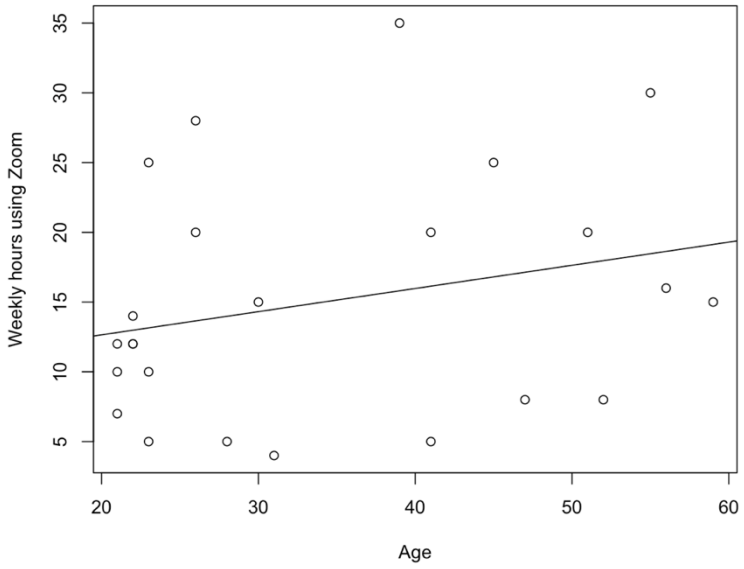

Figure 6: A scatterplot of the effect of age on the average amount of hours using Zoom per week. *The correlation is insignificant.*

To see whether age affected the subjective NASA TLX scores, a correlation test was executed between the NASA TLX scores and age. The assumption of normality wasn't met, so a Spearman correlation was executed. No significant correlation was found (S = 20659, p = 0.41). The data is shown in Figure 7. Note that every participant is shown twice in the graph, due to the fact that each participant filled in the NASA TLX after both the CoVince and the Zoom condition.

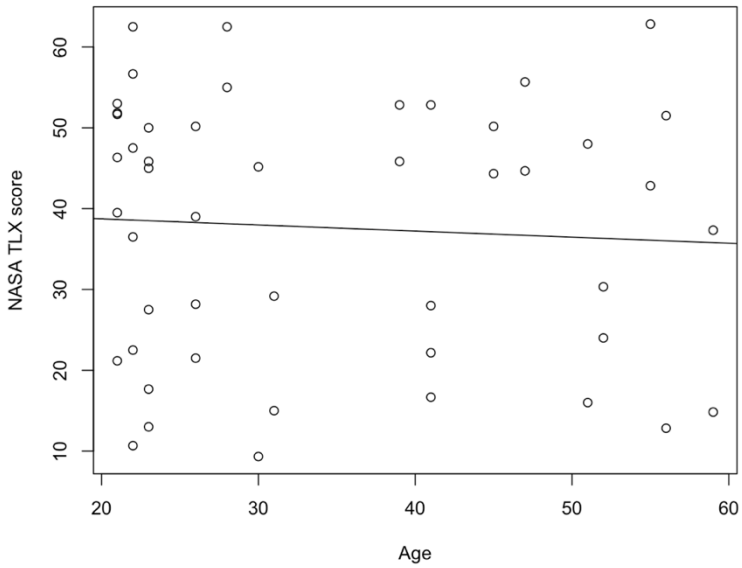

Figure 7: A scatterplot of the effect of age on the NASA TLX score given*. The correlation is insignificant.*

A correlation test between the hours spent using one of the software programs and the NASA TLX scores was executed to see whether experience with the software decreased the NASA TLX score. Because no participants used the CoVince software regularly, the correlation test was executed between the weekly hours spent using Zoom and the NASA TLX scores from the Zoom condition. No significant correlation was found ($r(22) = 0.53$, $p = 0.60$). The data is shown in Figure 8.

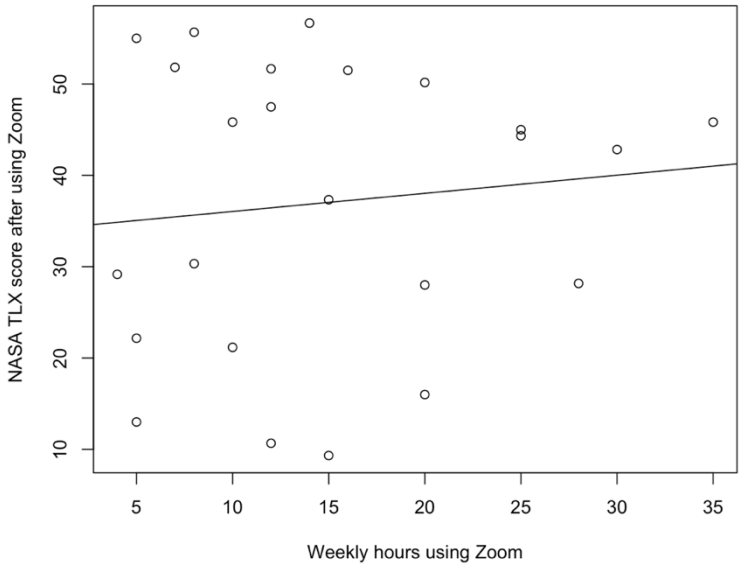

Figure 8: The effect of the average amount of hours using Zoom per week on the NASA TLX score given after having been in the Zoom condition. *The correlation is insignificant.*