A full shift field study to evaluate user- and process-oriented aspects of smart glasses in automotive order picking processes

Nela Murauer[1], Nerina Pflanz[2]
[1,2]BMW Group, Max-Diamand-Straße 5, 80788 Munich, Germany
[1]nela.murauer@bmw.de, [2]nerina.pa.pflanz@bmw.de

Abstract. Traditional flow production is experiencing a dynamic change towards smart factories. In the context of Industry 4.0 new technologies for worker assistance can be implemented in assembly, logistics or maintenance. Using Augmented Reality (AR) in order picking processes can provide advantages like time and error reduction and subsequently cost decrease. But what is the impact on employees wearing smart glasses during a full shift? Conducting a field study on the shop floor of an automotive production plant with expert order pickers, we analyze health-oriented aspects as well as the task completion time and the error frequency for a full shift usage of smart glasses in order picking processes. We face challenges such as an interaction with the warehouse management system and a predetermined production rhythm to ensure reliable results for the 8-hour usage of smart glasses, which may point the way towards future digital manufacturing.

Keywords: Augmented Reality, smart glasses, order picking processes, logistics, full shift usage

1 Introduction

The forth industrial revolution, known as 'Industry 4.0', forces manufacturing companies to shape the dynamic change from traditional flow production based on Taylorism towards smart factories. An increased amount of automatization is especially interesting for production plants with high personnel costs. But also other motives, such as the decrease of error rates and an increased level of quality, supporting workers with assistive technologies or surveillance of real time data show the entire range of opportunities of digital manufacturing. Changing logistic processes we focus on autonomous transportation systems like forklifts and tugger trains, robotics for efficient handling of empties, for automatic sorting and for truck unloading. Additionally, we conduct research on virtual reality for logistic planners and worker support with the aid of Augmented Reality (AR) and mobile devices such as smart watches. There is potential to automate many processes alongside the entire value stream, but especially in order picking processes we profit from the employees' flexibility regarding handling different objects and materials [1]. Therefore we engage in the context of Industry 4.0 in an order picking environment with assistive technologies for workers instead of their replacement by robotic solutions. Manual man-to-goods-order picking processes, where the positions of pallet cages to pick from are fixed and the worker collects all required components by moving from one pallet cage to the other [2], are the main focus of our research project. One highly promising technology is the usage of AR visualizations in head-mounted displays (HMD) or smart glasses, which we regard as synonymous in this paper. By displaying order picking information in the worker's field of view, we hope to decrease task completion times due to the elimination of unnecessary head- and body-movements. Integrating conspicuous error feedback may provide a reduction of error frequency. Additionally, visual guidance of workers can be a tool for an increased amount of job rotation, which contributes to the flexibility of the shop floor management and motivates employees by learning new processes.

Many studies in the context of AR for manufacturing use cases, which are explained in detail in Chapter 2, are laboratory studies with young participants without prior order picking experience. Thus there are no reliable reference points for a full shift usage of smart glasses in an industrial context. A further point, which received no attention, is the interaction of AR-devices with a warehouse management system. Few studies consider interaction methods, for example a button attached to a belt (but none corresponding to common interaction methods in logistics departments) and most use a Wizard-of-Oz-technique. In addition, most of the studies excluded workers wearing corrective glasses from participating. As all of the points mentioned above are highly relevant for a trailblazing decision for or against the implementation of smart glasses in a full shift operation, we conduct a field study concerning the following research question: *What is the impact of a full shift usage of smart glasses in order picking processes on workers and the process?* Hereby, we choose a real workstation in an automotive production plant as test environment, we use experienced order pickers of all ages as participants, we do not exclude corrective glasses wearers and we interact in real time with the warehouse management system.

2 Related Work

2.1 Order picking processes

Especially in times of increased individualization and variant diversity, order picking is one of the most important departments of intra logistics [3]. In automotive industry order picking workstations serve as suppliers for the assembly line. At so-called 'supermarkets' the pickers put the components precisely in the same order of the further assembly into the movable target shelf [2]. The order picking method at theses supermarkets is a manual man-to-goods order picking, which means the position of the provided components are fixed and the picker moves between the pallet cages during the component collection. So the picker has to run through the whole warehouse in the worst case scenario [4]. The pallet cages to pick from are refilled by internal or external suppliers [5]. Once a target shelf is filled, a transportation system such as a tugger train brings the shelf to the assembly line.

There are different methods for the visualization of order information and for the interaction with the warehouse management system. Pick-by-Paper means the listing of orders on a printed paper, which does not allow any automated interaction with the warehouse management system. The picker ticks all the picked components on the paper list [6]. Disadvantages are a lack of error feedback, no possibility to work hands-free and a lack of real time interaction with the warehouse management system. Using Pick-by-Light requires a confirmation with the aid of a confirmation button [4] attached to the withdrawal or target box. This method requires high investment costs, but allows hands-free work [2], real time interaction with the warehouse management system and error feedback. Pick-by-Voice is based on speech recognition and auditory orders. Hands-free work is possible, but this method can be demotivating for the picker due to auditory isolation of the environment [3] [2] [4]. Pick-by-Vision refers to worker support by AR-technology [7]. Required order picking information is visualized in the employee's field of view. Monocular or binocular head mounted displays can be used as hardware. The visualization of information can be context-sensitive, which means location-dependent signs like arrows or frames, or context-independent without tracking [8]. Advantages are the avoidance of unnecessary head- or body-movements, the enabling of hands-free work and a flexibility regarding provided information.

2.2 Augmented Reality in production

In our literature research we found several studies regarding AR-technologies in a productive environment. Büttner et al. [9] give an overview about AR- and virtual reality-related publications in a manufacturing context. Tümler [8] investigated AR-supported order picking processes with 20 participants with a highest age of 35 years in his laboratory study. He used a Microvision Nomad HMD and compared it with a paper-based order list. A forearm keyboard served as interaction device. In his study Tümler focused on process- and user-oriented variables. Among others, he analyzed eye dominance, task completion time (TCT), error frequency, heart rate variability and subjective strain. As a result an increase of headache and subjective workload was noticeable for the AR-system in comparison to Pick-by-Paper. Additionally, he observed four times more type errors and 30% less picked components using AR, witch means a higher task completion time for this visualization method. Summarizing his results, he expects other results for experienced pickers. Furthermore he postulates studies for an 8-hour usage. Based on the amount of omission errors, which increased during the test period, he recommends further research regarding concentration performance. Wiedenmaier [10] tested three different AR-assembly use cases with 36 participants with a highest age of 31. He also used a Microvison Nomad HMD and compared the usage with a paper-based scenario. The participants interacted with a mouse click with the HMD and did not get any error feedback. Whereas the error frequency is seen as task dependent, the subjective strain value showed no changes. Regarding the TCT, he determined time savings of 23% with the AR-system. Kampmeier et al. [11] focused in their laboratory study on ophthalmological variables. They tested paper-based instructions, paper-based instructions wearing a switched off HMD and a switched on HMD during 7.5 hours. 45 participants, mostly students with a highest age of 33, had to pick and to assemble technical toys. A portable PC worn at the belt served for the interaction with the HMD. Among others, they analyzed the d2-test of attention [12], the working quality and quantity and the heart rate variability. In their results, they observed a 13 percent increase in headaches wearing the HMD. Due to the fact that this increase was also detectable with a switched off HMD, they summarize that hardware-related reasons cause headaches. Nevertheless the participants classified the HMD wearing comfort as acceptable. Analyzing the different tasks, they regard AR-support as more suitable for picking than for assembly tasks. With reference to the ophthalmological results, they assume 'that from a medical and work psychological point of view there is nothing against a full shift usage of an AR-system'. Finally they recommend an improvement of the wearing comfort and a continuous monitoring of the users in case of a roll-out [11]. Büttner et al. [13] conducted a similar laboratory study. 13 unexperienced participants had to pick and to assemble LEGO ® animals. Focus of the study was the comparison of HMD-based, projection-based and paper-based instructions. As hardware served Vuzix Star 1200 glasses and a projector Optoma GT760. An interaction method is not explicitly mentioned. Referring to their results, the TCT during the HMD scenario was significantly higher than in both others. The error frequency behaves in a similar way. It is for the HMD scenario higher than in the projection- and the paper-based scenario. For the evaluation of the acceptance they asked the participants for a classification of the ease of use, the helpfulness and the joyfulness. Regarding the helpfulness and the joyfulness, they rated the projection- and paper-based scenarios significantly higher than the HMD-scenario. Thus they remarked that the 'combination of corrective glasses with HMD is not considered in most studies and settings' [13], and included spectacle wearers in their study. Guo et al.'s [14] study lead to conflicting conclusions. They investigated in a laboratory study the differences between Pick-by-Paper, Pick-by-Light, a cart-mounted display and a HMD (Microoptical SV3 opaque with laptop carried on the back). Eight unexperienced participants with an age of 22 till 27 years picked orders. Instead of the participant, the supervisor interacted with the system. Using the HMD the TCT was significantly lower than Pick-by-Light and Pick-by-Paper. In contrast to Büttner et al.'s observations the error frequency was lower in the HMD-scenario. Additionally, the subjective strain value was the lowest of all scenarios even though the hardware was uncomfortable. Reif and Günthner tested in their study the difference between Pick-by-Paper and Pick-by-Vision together with an industrial partner. Their 16 test candidates – students, researchers as well as skilled workers with an average age of 27.6 years – wore a HMD for a time period from 30 to 45 minutes. Speech input served as interaction method. As a result they observed a reduced task completion time by 4% using the AR-system in comparison with the paper list, which is not a significant difference. The error frequency was seven times higher using the paper list, but even this result does not lead to a significant result. Analyzing the subjective workload, the observations reinforce the effects with a lower subjective workload perceived after an AR-usage. Speech recognition as interaction method is viewed skeptically by the participants. In addition, they miss having an overview of all components to be picked. Consecutively, Reif and Günthner note that their results are not sufficient for a prediction of consequences of a full shift usage of an AR-system [15].

3 Methodology

In our field study, we compare two different visualization devices for displaying order picking tasks. The first device, which was implemented shortly before our study, is a cart-mounted monitor. The second device are binocular smart glasses. Conducting our study we analyze the differences of the impact on workers and the process using each device during a full shift.

3.1 Experimental environment

To find the best workstation as test environment, we analyzed all order picking workstations, so-called 'supermarkets', of the direct assembly delivery department at our production plant using Rasmussen's skill-rules-knowledge framework [16]. With Rasmussen's framework [17] human activities can be subdivided into skill-based, rule-based and knowledge-based tasks. According to [10] especially rule-based activities are suitable for an AR-usage. So we analyzed every single process step of all workstations and selected the workstation with the most rule-based process steps as test environment. The chosen workstation is for two workers. In a predetermined rhythm, analogous to the assembly line, they pick footwell claddings and put them into the target shelf in the same order as processed at the assembly line. There are different variants of footwell claddings (different colors, model series, driver's sides and components for emergency vehicles), which are located in a u-shaped layout at the workstation. Variant A and B were abolished a few month ago, so variant C is the first and variant L the last. Every variant position consists of two pallet cages, placed on the left and on the right of the way, providing components for the left side of the car and for the right side of the car. Figure 1 shows a picture of the chosen supermarket with the component designations and their color coding.

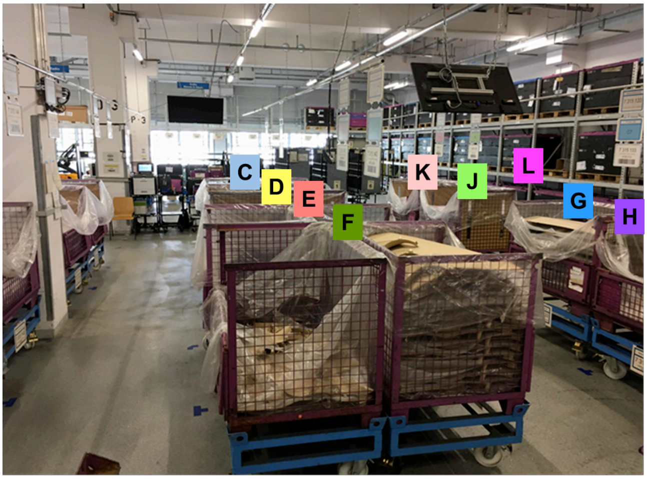

Fig. 1. Test workstation 'supermarket footwell claddings' with components C, D, E, F, G, H, J, K and L

The picker moves from C to L, picks all components of one variant and puts them into the target bins in the movable shelf. Six movable shelves are circulating. Three are at the workstation and three others on the way to the assembly line, at the assembly line or on the way back to the workstation. A tugger train brings the full shelves to the assembly line and brings back the empty ones. The target shelf consists of 16 consecutively numbered target bins. If the order contains components of the variants C, G and J, he puts C-components for example in bin 1, 3, 5, 7, 9, 10, 15, G-components in bin 2, 4, 6, 11 and J-components in bin 12, 13, 14 and 16. In doing so, he scans the pallet cage, from where he picked, and the target bin, which he filled, to interact with the warehouse management system. Additionally to the picking process, one pre-assembly has to be made for every pair of footwell claddings. Depending on the bought special equipment a boot switch or a shutter has to be assembled on the driver's side of the car. Furthermore, five white clips per pair must be fixed. While one worker picks the components, the other assembles the white clips. The pickers change role every other shelf, so that one picks and one clips during one round.

3.2 Hardware

The actual monitor is an innovation at this workstation and was implemented shortly before our study. The monitor is fixed on the right side of the movable shelf and can be repositioned easily from one rack to the next by the picker. Before this implementation pickers read the orders on a monitor, which hang from the ceiling.

The implementation of the new monitor was accompanied by the connection to the SAP warehouse management system. Scanning barcodes serve as interaction mechanism with the system. As scan device we use a scan glove, called 'ProGlove' of a Munich-based startup. Compared to classical hand-held barcode scanners, workers can work hands-free and the weight of the scanner is much lower. The scanner itself is changeable just like the cuff. A button in the palm of the hand triggers the scanner. Figure 2 shows the ProGlove cuff and the scanner.

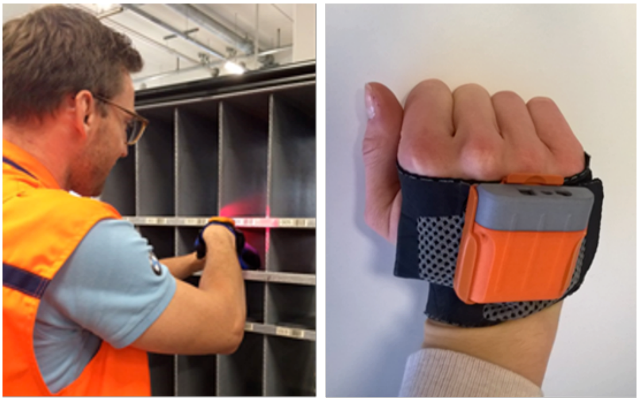

Fig. 2. Scan-device 'ProGlove' [18]

As visualization device we use ODG R7 binocular smart glasses. They have a 30-degree Field-of-View, weigh 180g (compared to the 579g heavy Microsoft Hololens) and they are usable for persons wearing corrective glasses by fixing optical lenses behind the front glases on the nose clip.

In a pre-study we compared different scan-mechanisms using smart glasses as vizualization device [18]. First the participants used the glasses also as scan-device, then we tested the combination of smart glasses as visualization tool and a ProGlove as scanner. We used the same target shelf and processes as at the real workstation, where we test now, but with the difference that we provided three instead of nine variants. Even tough the error frequency was the lowest in case of scannning with the glasses, the mean task complection time was higher than in the actual process (8:11min; SD ±01:08min, monitor: 5:52min; SD ±00:31min). Using the ODG-ProGlove-combination was faster than both other scenarios (4:57min, SD ± 00:29min). Both, subjective workload and the usability assessments, led to worse results for using the glasses as scanner than for the combination or the actual monitor. Based on this results we decided to use an external ProGlove scanner in combination with ODG R7 smart glasses as visualization device for our main study. For detailed information about our pre-study see [18].

3.3 Visualization

For a user-centered interface design, we conducted a design-thinking workshop with future users directly on the shop floor. 40 participants had the opportunity to design and evaluate their desired solution. Based on paper prototyping, which is a common method in app design, we adapted this method for an AR-context. For further information about the new methodology called 'Photo Prototyping' and a new process oriented idea collection tool, called 'Morphological Storyboard', see [19].

The preferred solution is context-independent and contains the display of the boot switch (pre-assembly) on the left side, then the component name, followed by a color-coded square and a series of target bin numbers. Once the placement of a component is confirmed by a scan, the first target bin number disappears and the remaining numbers move to the left. Additionally, the pickers asked for error feedback, thus we added a brief illumination of a green check mark or a red ‘x’. An example of a visualization shows figure 3:

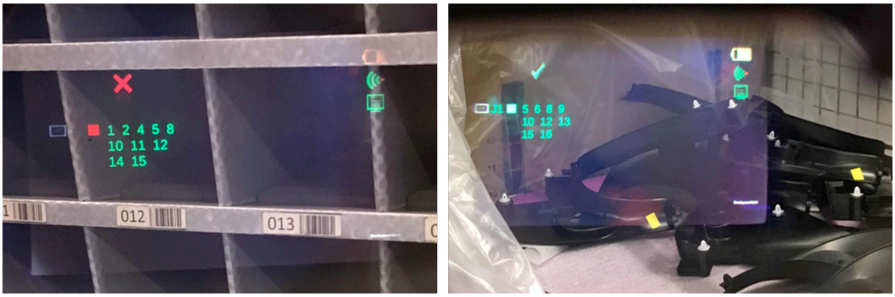

Fig. 3. Error feedback in the ODG R7 smart glasses

3.4 Study design

In our main study, presented in this paper, we focus on the impact of a full shift usage of smart glasses on the workers and the process. As explained above, an order picking workstation for footwell claddings at our automotive production plant in Munich serves as test environment. Inspired by [20] we tried to motivate the employees to participate in the study by providing sandwiches, sweets and coffee during one full shift. At our coffee-station the workers were open to talk and participate in the design thinking-workshop. At the same time they could volunteer for the main study. As a reward for the participation we announced a team event for all participants. All 23 participants are real logistics employees (20 male, 3 female). For the three spectacle wearers, we ordered individual manufactured corrective lenses for the smart glasses. Before the start, we had to discuss every variable to evaluate with the general works council and the data protection department. To avoid that they exercise their right of a veto, we accommodated their wishes. For this reason, we do not survey the age of participants and do not analyze personal differences between the scenarios but the overall average for each scenario. However, participants range from young workers with merely 2 years of working experience to long-employed workers who will retire in only 2 years.

We conduct our study during the morning shift, which starts at 5:50 a.m. and is finished at 14:55 p.m. During the shift, there are two breaks of 15min and one break of 30min. Consequently, the net working time is eight hours and five minutes. The process of the study is in both scenarios identical. We decided against a randomized design for two scenarios, which means we first test the monitor and then the glasses. Both scenarios are new for the pickers. One reason for our decision is that only the impact of the real scenario (old system first and then the innovation) counts for manufacturing companies, including all learn effects. Another point is that we have to change the SAP - warehouse management system to test the smart glasses. For a randomized design, we would have been forced to make this very change every day, which poses a risk. As we test at a real workstation, which delivers to the assembly line in a predetermined rhythm, we decided to avoid the risk of an assembly line stop caused by software changes for our workstation.

At the beginning of the first test scenario we ask the participants to rate their order picking experience. Additionally, we tested their ocular dominance. Before we start the first order picking task, we conduct the d2-Test of Attention [12], the Simulator Sickness Questionnaire (SSQ) [21] and the Visual Fatigue Questionnaire (VFQ) [22]. During the shift, we stop the task completion time (TCT) per shelf, we count the error frequency and we ask the participants to evaluate their mood six times per shift using a smiley scale, inspired by [23]. After the shift we repeat the d2-Test of Attention, we add the Raw NASA-TLX [24], we repeat the SSQ and the VFQ. To get further comments from the participants, we do a brief guided interview at the end of the shift. Table 1 gives an overview about the timeline and the methodology. Method details are briefly explained in our pre-study [18] and will be explained below.

Table 1. Methodology of the full shift study

Pre-test	Test	Post-test
Ocular dominance test	Error frequency	d2-Test
d2-Test		
SSQ	(type error, omission error,	Raw NASA-TLX
VFQ	scan-error)	SSQ
	Task completion time	VFQ
	Smiley scale	Guided interview

3.5 User-oriented variables

The d2-test of attention is a common tool for the evaluation of concentration performance. It consists of 14 lines with ‘d’s and ‘p’s. Below and above the letters one to four marks are added. During 20 seconds per line the participant has to mark as much ‘d’s with two marks (two below, two above or one below and one above) as possible. With the aid of a stencil omission errors and confusion errors can be detected. The concentration capacity ‘CP’ can be calculated by subtraction the confusion errors from the amount of right detected ‘d’s with two marks per line. Conducting this test we adhere to the guidelines of the test-manual [12].

The Simulator Sickness Questionnaire is a qualitative questionnaire, which contains 16 variables of subjective discomfort: *general discomfort, fatigue, headache, eyestrain, difficulty focusing, increased salivation, sweating, nausea, difficulty concentrating, fullness of head, blurred vision, dizzy (eyes open), dizzy (eyes closed), vertigo, stomach awareness* and *burping* [21]. The evaluation scale is word-based and contains four steps: none, slight, moderate, severe, scored with 0, 1, 2 and 3. Based on a factor analysis [21] found three symptom groups ‘nausea’, ‘oculomotor’ and ‘disorientation’. Value N is influenced by the individual scores of *general discomfort, increased salivation, sweating, nausea, difficulty concentrating, stomach awareness* and *burping.* The sum of the individual scores must be multiplied by 9.54. Value O is based on the individual scores of *general discomfort, fatigue, headache, eyestrain, difficulty focusing, difficulty concentrating and blurred vision.* The multiplying factor is 7.58. Value D contains the individual scores of *difficulty focusing, nausea, fullness of head blurred vision, dizzy (eyes open), dizzy (eyes closed)* and *vertigo.* Value D is multiplied by 13.92. The total score (TS) is calculated as sum of N, O and D, multiplied by 3.74 [21]. For our study we use the German translation of [25].

To evaluate the perceived change of eye-related variables we use the Visual Fatigue Questionnaire. 17 items (*dry eyes; watery eyes; eyes are irritated, gritty, or burning; pain in or around the eyeball; heaviness of the eyes; problems with line-tracking; difficulty in focusing; ‘shivering/jumping’ text; ‘foggy’ letters; glare from lights; blurry vision; double vision; headache; neck pain; dizziness; nausea; mental fatigue)* have to be evaluated on a Likert-scale from ‘not noticeable at all’ to 'extremely noticeable' [22]. In contrast to [26], we maintained the scale from 0-8 instead of expanding it from 0-10. Due to the absence of a total value, all items have to be analyzed individually [26]. For our future analysis we name the items a-q. A German translation is provided by [26].

For the analysis of user-oriented variables during the picking process, we created a smiley scale, inspired by [23]. The six step Likert scale contains smiley faces from very bad mood (6) to very good mood (1). The scores 6 to 1 are based on the German school marks. We ask the participants to evaluate their general mood six times per shift to get a course of the inter-daily mood development.

For the measurement of the subjective workload we use the Raw Nasa-TLX. The questionnaire measures six items *mental demand, physical demand, temporal demand, performance, effort* and *frustration* using a Likert scale from 0 till 100 [24]. The mean value of the six items equate to the final score. [27] and [28] provide German translations.

3.6 Process-oriented variables

During the shift we stop the task completion time (TCT) and count the error frequency. To get more detailed information we subdivide errors into error types. In literature we found a subdivision into quantity errors, type errors, omission errors and quality errors [29] or a subdivision into 'wrong amount, [...] wrong item [...], missing article [...] [and] damaged article' [30], which are usable as synonymous. Due to a lack of influence of the visualization device on the quality of components, we do not survey this error type. Additionally we do not count quantity errors. The target boxes of the shelf at our test workstation are too narrow to put more components into the bin, so that quantity errors can only be omission errors [18]. Summarizing we focus on omission and type errors. Furthermore we add scan-errors to our evaluation, which means that the picker has correctly picked and put, but scanned a wrong barcode. To measure corrections of the worker during the process, we subdivided every error type into 'corrected' and 'uncorrected' errors, which means self-discovered and corrected errors and errors, which will be detected at assembly line. Especially the uncorrected errors are extremely relevant in the industrial context. In case of an error detection the foreman has to drive as fast as possible to the assembly line for changing the component. If that is not possible within the assembly rhythm, this error will cause reworking, which contains at worst a step-by-step disassembly and reassembly. For stopping the TCT, we defined the process start at the moment, when the picker touches the shelf to start picking and we defined the placement of the shelf on the parking position as process finish. To avoid interfering influences, we make sure that enough components are in the pallet cages. If a pallet cage has to be changed, we ask the forklift driver to do this between two picking processes.

3.7 Hypotheses

Based on literature research and the results our pre-study, we formulated hypotheses for the main study comparing the two visualization devices monitor (M) and smart glasses (AR). The formulation of hypotheses regarding the positive or negative impact of study performance on workers and the process is based on the indicators extracted from preliminary results in a pre-study or suggested impacts in scientific literature.

H1: The mean task completion time changes significantly regarding scenario M in comparison to scenario AR.
H2a: The mean amount of corrected type errors per shelf changes significantly regarding scenario M in comparison to scenario AR.
H2b: The mean amount of uncorrected type errors per shelf changes significantly regarding scenario M in comparison to scenario AR.
H3a: The mean amount of corrected omission errors per shelf changes significantly regarding scenario M in comparison to scenario AR.
H3b: The mean amount of uncorrected omission errors per shelf changes significantly regarding scenario M in comparison to scenario AR.
H4: The mean amount of scan-errors per shelf does not significantly change regarding scenario M in comparison to scenario AR.
H5: The inter-daily difference of the mean smiley rating does not significantly change regarding scenario M in comparison to scenario AR.
H6: The amount of the mean raw NASA-TLX does not significantly change regarding scenario M in comparison to scenario AR.
H7: The inter-daily difference of the mean concentration performance does not significantly change regarding scenario M in comparison to scenario AR.
H8: The inter-daily difference of the mean Total SSQ-Score changes significantly regarding scenario M in comparison to scenario AR.
H9a-9q: The inter-daily difference of the value of each item of the VFQ changes significantly regarding scenario M in comparison to scenario AR.

4 Results

We conducted our study with 23 participants. Figure 4 shows their picking experience. Two participants have less than 6 month experience, three 6 month - 2 years, three 2-3 years, six 3-5 years, seven 5-10 years and two more than 10 years.

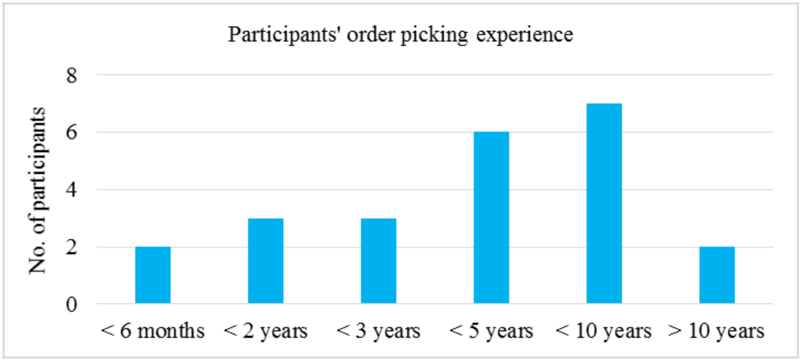

Fig. 4: Participants’ order picking experience

4.1 Process-oriented variables

For the statistical evaluation of the TCT and the error frequencies, we conducted a two sided t-test in a paired sample (α=0.05).

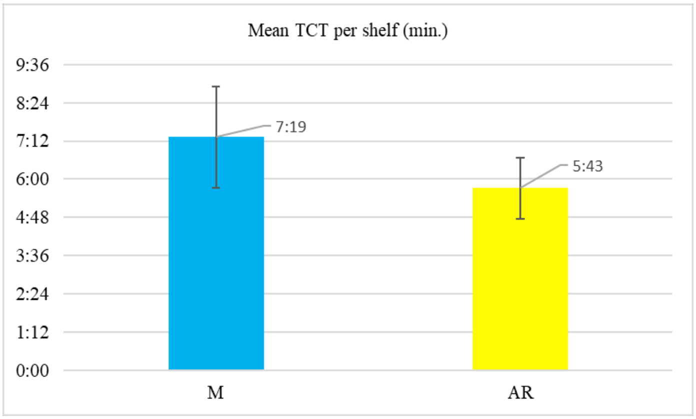

Fig. 5. Mean TCT per shelf for scenarios M and AR

Comparing the mean TCT per shelf, we observed a highly significant difference (p= 0.00) between the two scenarios, see Figure 5. While the mean TCT in scenario M was 7:19min/shelf (SD ±1:35min), we measured 5:43min/shelf for scenario AR (SD ±0:58min/shelf). The difference amounts to an average time reduction of 1:36min/shelf and 22%. Subsequently we maintain hypothesis H1.

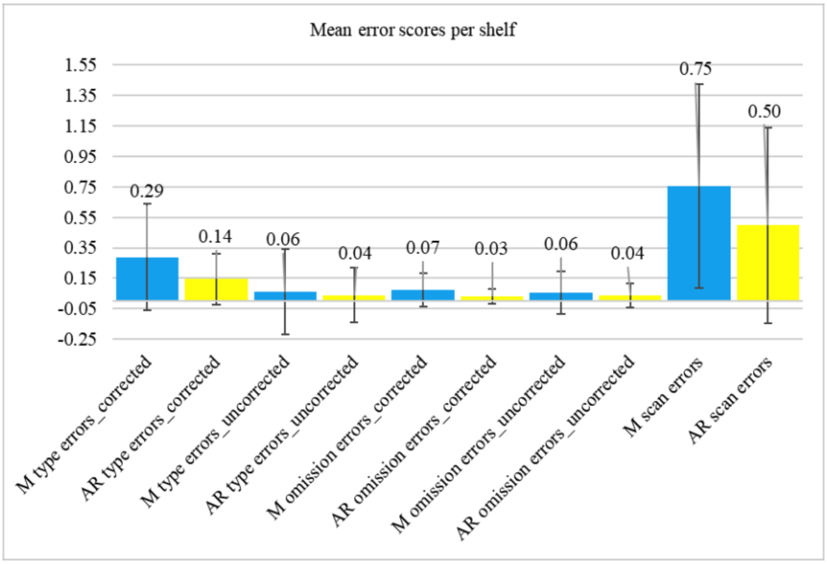

Fig. 6. Average error frequency per shelf for scenarios M and AR

Analyzing the error frequency in the predetermined error categories, we observed a decrease of the average amount of corrected type errors from 0.29 errors/shelf (SD ±0.35) to 0.14 errors/shelf (SD ±0.17) (p=0.05), of uncorrected type errors from 0.06 errors/shelf (SD ±0.28) to 0.04 errors/shelf (SD ±0.18) (p=0.25), of corrected omission errors from 0.07 errors/shelf (SD ±0.11) to 0.03 errors/shelf (SD ±0.05) (p=0.09), of uncorrected omission errors/shelf from 0.06 (SD ±0,14) to 0.04 (SD

±0.08) (p=0.56) and of scan errors from 0.75 errors/shelf (SD ±0.67) to 0.50 errors/shelf (SD ±0.64) (p=0.06) regarding scenario M in comparison with scenario AR. Subsequently the difference between the average amount of corrected type errors in scenario M and AR is significant. Therefore we maintain H2a. All other differences are not significant, thus we reject hypothesis H2b, H3a, H3b and maintain H4. Figure 6 shows an overview of the results.

4.2 User-oriented variables

A six step smiley rating scale serves for the detection of mood changes during the full-shift order picking operation. Figure 7 shows the two mean courses of the rating for scenario M and AR.

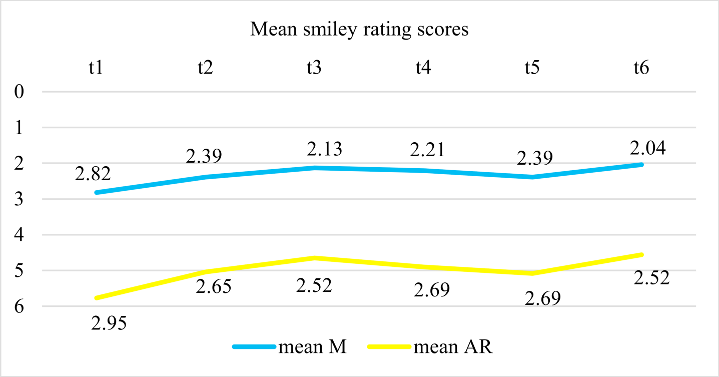

Fig. 7. Mean course of the smiley rating curve for scenario M and AR

For both scenarios the participants' mood got better during the morning shift. It is conspicuous that the course of scenario AR already starts worse than the one of scenario M. Both behave nearly parallel. From the beginning of the shift to the first break (t3) the curve develops positively, then until the 30min-lunch break (t5) it decreases and increases until the end of the shift. The curve during the AR-usage does not show any abnormalities in comparison to curve M. In addition, we could not determine any influence of working with smart glasses on the mood. The highest (worst) value of both scenarios is the baseline measurement before the shift started. Due to the lack of a significant difference of the mean inter-daily smiley rating (α=0.05) (p= 0.57) regarding scenarios M and AR, we maintain hypothesis H5.

With the aid of the Raw NASA-TLX questionnaire we statistically analyzed the perceived workload after the full shift usage of the visualization devices. Figure 8 shows the results.

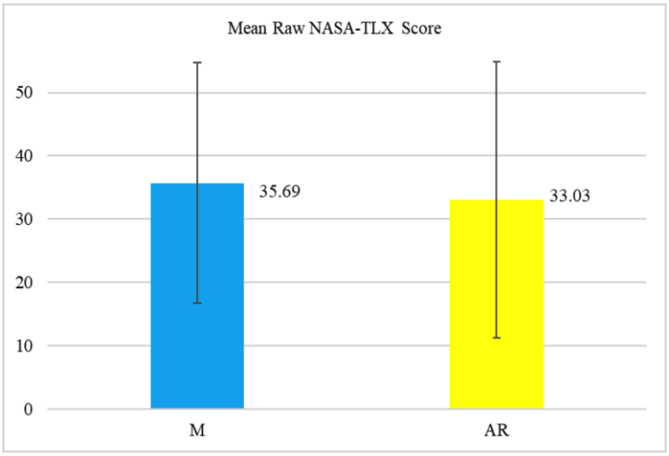

Fig. 8. Mean Raw Nasa-TLX score for M and AR

Regarding the mean raw NASA-TLX of the two scenarios, the difference is not significant (α=0.05) (p=0.53) with 35.69 (SD ±19.00) for M and 33.03 (SD ±21.80) for AR. Subsequently we maintain hypothesis H6. Interesting findings serves the analysis of the six subcategories. Whilst the mean score of both scenarios is relatively similar, the values of the subcategories are clearly different. After using the monitor (M) the participants evaluated their physical demand with 40.87, temporal demand with 36.30 and effort with 39.78 higher than after the usage of the smart glasses (AR) (31.52; 26.52; 32.39). The mean values for mental demand, performance and frustration behave inversely with (44.57; 25.86; 26.74) for M and (47.83; 28.48; 31.52) for AR.

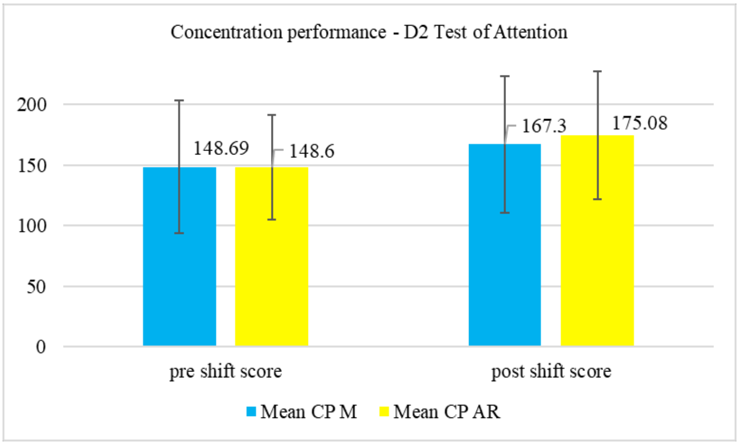

Fig. 9. Mean concentration performance (pre- and post-shift) for M and AR

Beside the NASA-TLX Brickenkamp's d2-Test of Attention provides interesting insights. Figure 9 presents the mean concentration performance values (CP) for both visualization devices M and AR and the inter-daily difference (pre-shift and post-shift) (α=0.05). Contrary to our initial expectations, the concentration performance increases during a full shift. The pre shift score of both scenarios is with 148.69 (M) and 148.6 (AR) similar. After using the monitor (M) the mean CP increases significantly by 18.6 to 167.3 (p=0.00). This enhances regarding scenario AR, in which the mean CP increases significantly by 26.48 to 175.08 (p=0.00). Comparing the mean inter-daily changes of both scenarios, we did not found any significance (p=0.26). Subsequently we maintain hypothesis H7. We assume that the early start of the morning shift influences lower scores in the morning and a gradual increase in concentration performance towards the end of the shift.

It is conspicuous that the standard deviation is quite high in both scenarios. This is probably based on a wide range of skill and educational levels in our sample.

Analyzing the mean pre- (t1) and post-shift (t2) scores of the Simulator Sickness Questionnaire brings interesting findings. The mean Nausea score starts with 15.35 (M) and 20.74 (AR) and increases to 25.72 (M) and 25.72 (AR). The inter-daily differences are significant (α=0.05) (p=0.04) for M and not significant (p=0.50) for AR. The oculomotor pre-shift values are with 33.29 (M) and 32.20 (AR) nearly similar and behaves similarly regarding their increase to 45.81 (M) and 46.14 (AR). Both inter-daily changes are not significant (p=0.11) (M) and (p=0.08) (AR). Disorientation increases during the working shift as well. For M the mean score nearly doubles from 21.16 to 41.15, which describes a significant inter-daily change (p=0.04). For the AR-visualization the value behaves similarly and changes from 26.02 to 49.63, which is not significant (p=0.07). Figure 10 gives an overview of the results.

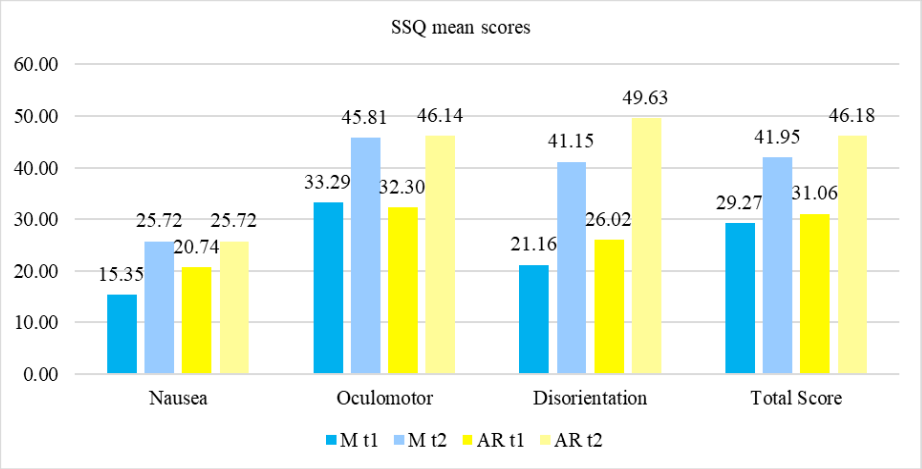

Fig. 10. Mean values of Nausea, Oculomotor, Disorientation and the Total Score for M and AR

The evaluation of the total score TS does not present any significant inter-daily changes (p=0.06) (M) and (p=0.12) (AR) with pre-shift scores of 29.27 (M) and 31.06 (AR) and post-shift scores of 41.95 (M) and 46.18 (AR). Regarding the mean inter-daily changes of the monitor-based picking in comparison with the smart glasses-based picking, there are no significances (Nausea p=0.49, the Oculomotor p=0.89, Disorientation p=0.75 and Total Score p=0.81). Subsequently we reject hypothesis H8.

The Visual Fatigue Questionnaire contains 17 items (*dry eyes (a); watery eyes (b); eyes are irritated, gritty, or burning (c); pain in or around the eyeball (d); heaviness of the eyes (e); problems with line-tracking (f); difficulty in focusing (g); 'shivering/jumping' text (h); 'foggy' letters (i); glare from lights (j); blurry vision (k); double vision (l); headache (m); neck pain (n); dizziness (o); nausea (p); mental fatigue (q))*. The individual categories of the used Visual Fatigue Questionnaire are reliable indicators on their own. The questionnaire is not designed to provide a single quantitative sum value as an indicator of visual fatigue.

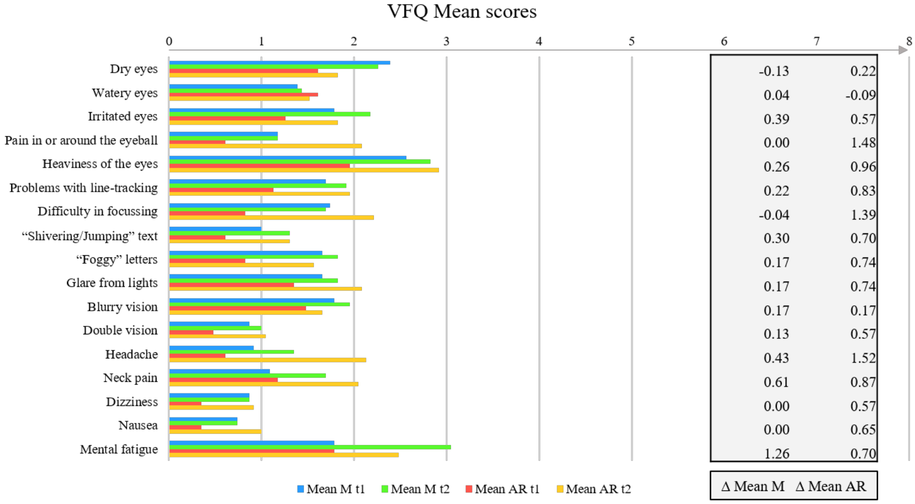

Fig. 11. Mean values of VFQ-items a-q for M and AR

Figure 11 presents the mean inter-daily deltas of the individual VFQ-scores of both scenarios M and AR. Analyzing the development of the scores caused by the monitor-based picking the scores range from an increase by 1.26 to a decrease of 0.13 for scenario M in comparison to an increase of 1.52 to a decrease of 0.09 for scenario AR. Only two items, watery eyes (b) and mental fatigue (q), worsen more working with the monitor than with the AR-system during the day. Item k, blurry vision, does not change depending on the visualization device. All other items increase more regarding the inter-daily difference of scenario AR. It is obvious that our participants experienced the most negative development during the shift for the items m 'headache', d 'pain in or around the eyeball' and g 'difficulty in focusing'. On a scale from zero to eight the mean value for item m increases by 1.52 (AR), item d by 1.48 and item g by 1.39. For said items we observed significant differences between the mean inter-daily changes of both scenarios (α=0.05) (item m: p=0.0217; item d: p=0.0068; item g: p=0.0214). Further significances could not be established in our sample. Summarizing, we merely maintain hypotheses H9d, H9g and H9m and reject hypotheses H9a, H9b, H9c, H9e, H9f, H9h, H9i, H9j, H9k, H9l, H9n, H9o, H9p and H9q.

4.2 Guided interviews

For getting additional qualitative feedback, we asked the participants for their perceived positive and negative aspects of the full shift AR-usage. Additionally, we motivated participants to voice suggestions for improvement or for suitable use cases and workstations where they could imagine wearing smart glasses during a full shift operation. Summing up, the pickers' opinions about the full shift usage of smart glasses is combination with a ProGlove as interaction device are quite different. Some employees 'had fun' and think that working with the glasses is 'cool'. Most of them liked the user interface and the colors, especially the series of numbers, with which they had a better overview. Providing for error feedback is appreciated and leads to 'a nearly error-free picking'. The higher working speed is viewed as advantageous, caused by a visualization in their field-of-view and an avoidance of head- and body movements. Wireless working without any power bank prevents entanglement. Spectacle wearers liked the corrective clips and did not notice the weight difference between smart glasses and spectacles. One picker described the tasks as 'robot work, where we do not think, but only act', which he liked. Negative aspects were mostly hardware related comments. Most pickers perceived the weight of the glasses as uncomfortable and did not like the imprint on their nose caused by the nose clips, with which they did not want to go to the break room. Some participants found the temples inflexible and narrow. In fact, they saw a connection between the headache at the back of the head and the design of the temples. Another challenge was to refocus from objects to visualization in the glasses and the limited field of view. Few pickers perceived the display as blinding.

Suitable workstations from the participants' perspective are logistics, especially order picking workstations and pre-assembly, but not the assembly line. Some pickers firmed that the test workstation is a suitable working environment for an AR-support. Another aspect are training scenarios, in which AR-guidance leads unskilled workers through the process. For a future 8-hour usage, we gathered many proposals. Most suggestions concerned the wearing comfort of the smart glasses: lower weight, individually adjustable temples, padded nose clips and easy adjusting of contrast and brightness. Few ideas for improvement pertain to the design of the user interface, with which most pickers were satisfied. The symbols for battery status and wifi connection are not necessary in the field of view and can be replaced by attached lights on the side of the glasses. Some workers wished the numbers were bigger and located further down in the field of view.

5 Discussion & Conclusion

Due to the lack of studies of AR-usage during a full shift operation in real industrial working environment, connected with real-time interaction with a warehouse management system and real order pickers of all ages as test candidates we conducted a field study at our automotive production plant in Munich. In our research we focused on exploring the impact of a full shift usage of smart glasses on workers and the process. Under the conditions of our testing scenario the AR-support contributed a 22% decrease of the mean task completion time. Depending on the error type, we observed a reduction of the mean error frequency up to 58%. The statistical analysis of the subjective workload did not yield significant differences between the scenarios M (monitor) and AR. Whilst the sub-values 'mental demand', 'performance' and 'frustration' of the NASA-TLX are higher after using smart glasses as visualization device, the sub-values 'physical demand', 'temporal demand' and 'effort' were higher using the monitor. We hypothesize that a higher physical demand for M is caused by head- and body-movements, which can be avoided though smart glasses. Due to the decrease of the TCT the temporal demand is lower for AR, but nevertheless the participants evaluate their frustration scores higher. Even though the objective variables such as TCT and error frequency point to better performance using AR, workers underestimate their performance. This effect is probably caused by a perceived uncertainty during working with a novel technology, with which they have never been in contact. An inter-daily increase of the mean concentration performance can stem from two possible reasons: either the time of day or the continuous mental demand using AR, which could support the workers' attention. Further research should investigate the effects of different work shifts such as a morning shift, an evening shift and a night shift. Regarding the results of the Simulator Sickness Questionnaire the values for scenario AR increase more than for scenario M, but the maxima are quite similar. Merely the end-of-shift AR score for disorientation is striking and indicates higher disorientation than after working a full shift with the monitor. Yet, no participant opted out of the study and or reported significant nausea. In fact, performance scores such as TCT improved during the day despite of prolonged use of the hardware. VFQ answers suggest that differences between afternoon and baseline measurements in the morning are higher for AR. These subjective assessments of the workload and effect of working a full shift with AR technology stand in contrast to Kampmeier et al.'s [11] findings. We cannot determine for certain whether the differences between our results and [11] were caused by a different testing environment, diverging test methods or another working task. Further insight into the topic would be welcomed.

Summarizing, the process-oriented variables are indicative of the suitability of AR technology for full shift usage in automotive order picking. Improving the wearing comfort of the smart glasses would support of a user-oriented approach to introducing such innovations to a productive working environment. Neither interacting with the warehouse management system nor the battery performance were an issue in conducting the field study. We did not observe that age or wearing corrective glasses hindered participants in working with smart glasses. We can imagine to introduce the tested technology at predetermined and well-suited workstations to gather further insights and requirements for long-term usage in the industrial environment.

References

1. Stinson, M. and Wehking, K.-H.: Leistungsbewertung und -optimierung in der manuellen Kommissionierung. In Proceedings Logistics Journal, Vol.2012., pp. 1-7, (2012)
2. Klug, F.: Logistikmanagement in der Automobilindustrie; Grundlagen der Logistik im Automobilbau. Springer, Heidelberg Dornrecht New York, (2010)
3. Reif, R.: Entwicklung und Evaluierung eines Augmented Reality unterstützen Kommissioniersystems. PhD Dissertation. Lehrstuhl für Fördertechnik, Materialfluss und Logistik, Technische Universität München, Garching (2009)
4. Martin, H.: Transport- und Lagerlogistik, Vieweg Teuber Verlag (2011)
5. Ohno, T.: Das Toyota Produktinssystem, New York Campus, Frankfurt (1993)
6. Hompel, M., Schmidt, T.: Warehouse Management: Automatisierung und Organisation von Lager- und Kommissioniersystemen. Springer Verlag, Berlin Heidelberg (2003)
7. Günthner, W. A.; Blomeyer, N.; Reif, R.; Schedlbauer, M.: Pick-by-Vision: Augmented Reality unterstützte Kommissionierung. Abschlussbericht zum Forschungsvorhaben, Garching, (2009)
8. Tümler, J.: Untersuchung zu nutzerbezogenen und technischen Aspekten beim Langzeiteinsatz mobiler AR-Systeme in industriellen Anwendungen. PhD Dissertation. Fakultät für Informatik, Otto-von-Guericke Universität Magdeburg, Magdeburg (2009)
9. Büttner, S., Mucha, H., Funk, M., Kosch, T., Aehnelt, M., Robert, S. and Röcker, C.: The design space of augmented and virtual reality applications for assistive environments in

manufacturing: a visual approach. Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments. ACM, (2017)

10. Wiedenmaier, S.: Unterstützung manueller Montage durch Augmented Reality-Technologien, Shaker, Aachen, (2004)
11. Kampmeier, J., Cucera, A., Fitzsche, L., Brau, H., Duthweiler, M., Lang, G.K.: Eignung monokularer Augmented Reality – Technologien in der Automobilproduktion. In: Klinische Monatsblätter für Augenheilkunde, Georg Thieme Verlag, Stuttgart, pp. 590-596 (2007)
12. Brickenkamp, R.: Test d2 Aufmerksamkeits-Belastungs-Test. Hogrefe Verlag für Psychologie, Göttingen Bern Toronto Seattle, (1962)
13. Büttner, S., Funk, M., Sand, O. and Röcker, C. Using Head-Mounted Displays and In-Situ Projection for Assistive Systems – A Comparison. Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments. New York, (2016)
14. Guo, A., Raghu, S., Xie, X., Ismail, S., Luo, X., Simoneau, J.& Starner, T.: A comparison of order picking assisted by head-up display (HUD), cart-mounted display (CMD), light, and paper pick list. In Proceedings of the 2014 ACM International Symposium on Wearable Computers, pp. 71-78, ACM (2014)
15. Reif, R. and Günthner, W.: Pick-by-Vision: augmented reality supported order picking, In: Vis Comput 25, pp. 461-467, Springer-Verlag (2009)
16. Murauer, N. and Gehrlicher, S.: Evaluation of order picking processes regarding the suitability of smart glasses-based assistance using Rasmussen's Skills-Rules-Knowledge framework. In Proceedings of the AHFE 2018 International Conference on Human Aspects of Advanced Manufacturing, July 21-25, 2018, Orlando, Florida, USA (2018)
17. Rasmussen, J.: Skills, Rules, and Knowledge; Signals, Signs, and Symbols, and Other Distinctions in Human Performance Models. IEEE Transactions on Systems, Man and Cybernetics, SMC-13, No. 3, pp. 257- 266 (1983)
18. Murauer, N., Pflanz, N and von Hassel, C.: Comparison of Scan-Mechanisms in Augmented Reality-Supported Order Picking Processes, Proceedings of the 6th Workshop on Interacting with Smart Objects (SmartObjects) in conjunction with CHI '18, pp. 69-76, CEUR Workshop Proceedings, http://ceur-ws.org/Vol-2082/paper_1.pdf , Montreal (2018)
19. Murauer, N.: Design Thinking: Using Photo Prototyping for a user-centered Interface Design for Pick-by-Vision Systems. Proceedings of the 11th ACM International Conference on PErvasive Technologies Related to Assistive Environments. ACM, New York (2018)
20. Rosenthal, R. and Rosnow, R.L. The Volunteer Subject. Wiley. New York (1975)
21. Kennedy, R. S., Lane, N. E., Berbaum, K. S. & Lilienthal M. G.: Simulator Sickness Questionnaire: An Enhanced Method for Quantifying Simulator Sickness, The International Journal of Aviation Psychology, 3:3, 203-220 (1993)
22. Bangor, A. W.: Display Technology and Ambient Illumination Influences on Visual Fatigue at VDT Workstations. Dissertation. Faculty of the Virginia Polytechnic Institute and State University, Blacksburg, Virginia, USA (2000)
23. Wong, D. L. and Baker, C. M.: Pain in Children: Comparison of Assessment Scales. In: Pediatric Nursing/ Jan. - Feb. 1988, Vol. 14, No. 1, pp. 9-17 (1988)
24. Hart, S. G. NASA-task load index (NASA-TLX); 20 years later. Proceedings of the human factors and ergonomics society annual meeting. Sage Publications, Vol. 50. No. 9., Sage CA, Los Angeles, CA (2006)
25. Neukum, A. and Grattenthaler, H.: Kinetose in der Fahrsimulation (Abschlussbericht Teil II, Projekt: Simulation von Einsatzfahrten im Auftrag des Präsidiums der Bayerischen Bereitschaftspolizei). Würzburg: Interdisziplinäres Zentrum für Verkehrswissenschaften an der Universität Würzburg. http://opus.bibliothek.uni-wuerzburg.de/volltexte/2013/7782/; URN: urn:nbn:de:bvb:20-opus-77829 (2006)
26. Wille, M.: Head-Mounted Displays – Bedingungen des sicheren und beanspruchungs-optimalen Einsatzes – psychische Beanspruchung beim Einsatz von HMDs. Bundesanstalt für Arbeitsschutz und Arbeitsmedizin, Dortmund (2016)
27. Niederl, T.: Untersuchungen zu kumulativen phychischen und physiologischen Effekten des fliegenden Personals auf der Kurzstrecke – Am Beispiel des Flugbetriebs der Boeing 737 Flotte der Deutschen Lufthansa AG, Dissertation, Institut für Arbeitswissenschaft der Universität Kassel, Kassel (2007)

28. Unema, P., Rötting, M., Sepher-Willeberg, M., Strümpfel, U. & Kopp, U. (1988). Der NASA Task Load Index: Erste Ergebnisse mit der deutschen Fassung. In Gesellschaft für Arbeitswissenschaft e.V. (Hrsg.). Jahresdokumentation 1988 der Gesellschaft für Arbeitswissenschaft e.V. - Bericht zum 34. Arbeitswissenschaftlichen Kongreß an der RWTH Aachen (S. 47). Köln: O. Schmidt.
29. Lolling, A.: Analyse der menschlichen Zuverlässigkeit bei Kommissioniertätigkeiten. Shaker Verlag. Herzogenrath, Germany (2003)
30. Schwerdtfeger, B., Reif, R., Günthner, W. A., Klinker, G., Hamacher, D., Schega, L., Böckelmann, I., Doil, F. and Tümler, J.: Pick-by-Vision: A First Stress Test. In: 8th IEEE International Symposium on Mixed and Augmented Reality 2009, Orlando, USA, pp.115-124 (2009)