Dietlind Helene Cymek, Dietrich Manzey

Sequential human redundancy: Can social loafing diminish the safety of double checks?

Citation details
Cymek, D. H., & Manzey, D. (2022). Sequential human redundancy: Can social loafing diminish the safety of double checks? In Journal of Experimental Psychology: Applied (Vol. 28, Issue 4, pp. 931–945). American Psychological Association (APA). https://doi.org/10.1037/xap0000439.

Sequential Human Redundancy: Can Social Loafing Diminish the Safety of Double Checks?

Dietlind Helene Cymek and Dietrich Manzey

Department of Psychology and Ergonomics, Technische Universität Berlin

Author Note

Correspondence concerning this article should be addressed to Dietlind Helene Cymek, Department of Psychology and Ergonomics, Technische Universität Berlin, Marchstr. 12, F7, 10587 Berlin, Germany. Email: dietlind.h.cymek@tu-berlin.de. The experiments were not pre-registered. Data, the analytic code to reproduce the results, and additional materials are available via the Open Science Framework (OSF) at https://osf.io/dga2n/?view_only=e783569673ef483f9d9521e83428a476.

Abstract

It is often assumed that if two people work on a failure-detection task one after the other, they will observe more failures than when only one person undertakes the task (four-eyes principle). However, human beings have also been found to exert less effort on tasks that they share responsibility for, a phenomenon called *social loafing*. In the current research, we assessed the effectiveness of sequential human redundancy in light of possible social loafing. In two laboratory experiments, teams of two participants performed a quality control task in a blinded and in a non-blinded condition, operationally defined by whether or not evaluations of the first checker were forwarded to the second one. In the blinded condition, no social loafing was found and a near-perfect overall team performance was observed. In contrast, non-blinded redundancy led to a substantial effort reduction of the second checker. However, despite this social-loafing effect at the second position, even non-blinded redundancy led to an overall safety advantage over a single-checker condition. Our research suggests that social loafing in sequential-human-redundancy work settings can occur but does not necessarily reduce gains in overall reliability. Blinded processes, however, seem to provoke less social loafing than non-blinded processes.

Keywords: group processes, motivation, reliability issues, process control, risk assessment

Public Significance Statement: Double checks are generally implemented to enhance safety. This research found that specific implementations of double-check processes may lead to reduced individual effort and may therefore limit the positive impact double checks theoretically might have.

SEQUENTIAL HUMAN REDUNDANCY: CAN SOCIAL LOAFING DIMINISH THE SAFETY OF DOUBLE CHECKS?

In high-risk industries, decisions and actions of outstanding importance, which could theoretically be made by a single person, are often taken by more than one individual. This sort of *human redundancy* (Clarke, 2005) is intended to raise the overall reliability of decisions, contributing to a reduction in failure, risk, or unnecessary costs as best as possible. We can find redundancy in many areas, and in various forms (Lerner, 1986). For example, on highly automated planes and on large ships, two or more people are usually present to monitor the autopilot system and to intervene in case the system malfunctions. Since everyone here is working on the task simultaneously, this form of redundancy has been termed *parallel redundancy*. Another form of redundancy present in the field is *sequential redundancy*. Here, the individuals involved do not necessarily carry out the task at the same time; indeed, by definition they must accomplish it in a sequential order. For example, two nurses independently determine medication doses to avoid incorrect dosages (Armitage, 2008), and two radiologists read the same X-rays to increase the sensitivity of breast cancer detection (Anttinen, Pamilo, Soiva, & Roiha, 1993). The latter two examples apply the *four-eye principle* or *double-check principle*, where an action or final decision can only occur after the contribution of two people. These principles are implemented in many fields, ranging from the assessment of important exams to the quality control of hazardous material. All these forms of redundant systems or backup systems are designed to ensure fail-safe operations. In case one component does not work completely reliably or is somehow biased (e.g., a radiologist misses a cyst during mammography screening), the other component may compensate this suboptimality by complementing or correcting its behavior.

The rise in the application of human redundancy grew out of the great success historically of technical redundancy. The implementation of redundant technical components, if independent and connected in parallel manner, led to major increases in system reliability in various applications, even if each individual component was only moderately reliable (Sagan, 2004). For example, let us assume

that an automobile has a split-circuit braking system. Each circuit can stop the car, and the circuits work independently so that even if one malfunctions, it does not compromise the other. If the reliability of each single brake circuit was 90.0% ([1-(1/10)]x100), their joint reliability would be 99.0% ([1-(1/10)^2]x100) so that the probability of the brakes failing is reduced from 10% to only 1% by adding the second circuit. Adding a third independent brake circuit would make the braking system even more reliable – 99.9% ([1-(1/10)^3]x100), turning a relatively unsafe system into a near-perfect one.

Systems harnessing human redundancy also often implicitly anticipate similar improvements. However, this assumption ignores the fact that human beings who work redundantly cannot be simply regarded as independent components. In most work settings that use human redundancy, the parties involved are usually aware of the redundant task fulfillment and their "backup role," so they may adapt their behavior and decision-making process in many (sometimes unconscious) ways. Based on an interview study in the medical sector, Armitage (2008) pointed out some specific risks of human redundancy. These range from the reduction of individually perceived responsibility and individual effort, to the neglect of redundant procedures under stress and lack of time, to the proneness towards a deference to authority.

The aforementioned risk of a reduction of individual responsibility and effort has already been investigated in the context of social loafing theory for some time (e.g. Kerr & Bruun, 1983; Latané, Williams, & Harkins, 1979; Shepperd, 1993). Social loafing is broadly described as the reduction of individual effort in group tasks compared to tasks with sole responsibility. It has been found for a variety of tasks, ranging from idea generation (Allport, 1920; Szymanski & Harkins, 1987), to clapping and shouting (Latané et al., 1979), to evaluating poems (Petty, Harkins, Williams, & Latane, 1977). All these studies showed that people exhibited a sizeable decrease in individual effort when performing in groups compared to when they performed alone. So far, social loafing has been mostly investigated in groups of two to sixteen individuals who perform a task simultaneously (parallel

redundancy). However, analogous effects may arise in pairs who perform safety-relevant tasks, such as double-check procedures, which follow a sequential logic. And if this is the case, it is important to know whether these social-loafing effects jeopardize the anticipated gains in reliability.

Social psychological theories and research on social loafing have described several factors that hinder or aggravate the potential loss of effort when people perform collectively, and thus might help to differentiate between relatively effective and ineffective human redundancy work settings. For example, Karau and Williams (1993) performed a comprehensive meta-analysis of 78 social-loafing studies and derived a variety of conditions that influenced social loafing. According to their analysis, important factors exist that contributed to social loafing in group settings. One factors is, for example, a *low evaluation potential of individual inputs*, that is most prevalent in group tasks where only the final team result is assessable, or were the individual inputs of two humans are hidden (or blinded to each other) to avoid mutual influence (Harkins, 1987; Harkins & Jackson, 1985; Harkins & Szymanski, 1989; Szymanski & Harkins, 1987). Another factor is *working with highly reliable coworkers*, since individuals may feel their input will have little impact on an already high-quality group product (Kerr & Bruun, 1983). In contrast, other factors such as the *meaningfulness of the task* and *group cohesiveness* were identified that mitigate social loafing. Karau and Williams (1993) integrated all these factors into what they dubbed the *Collective Effort Model* (CEM) and suggested that all these factors influenced individual motivation, by modifying the correlation between individual effort exerted and valued outcomes, given the general tendency of humans to try to maximize the utility of their actions. Interestingly, several studies that found a reduction of objective effort in group settings could not find a decrease in self-reported effort, suggesting that participants might not always be aware of their loafing, or might be hesitant to report it (Karau & Williams, 1993; Torka, Mazei, & Hüffmeier, 2021).

Given these factors and related effects, work settings using sequential redundancy can differ with respect to their proneness to social loafing effects, depending on how redundancy plans are

implemented and which of the two positions is considered. For example, in the case of mammography screening, work processes exist where radiologists are asked to withhold their diagnosis from one another before each of them has made a judgement; these processes are often referred to as being *blinded*. However, *non-blinded* work processes also exist, where radiologists second in line scrutinize X-rays in which calcifications or tumors have often already been marked by preceding radiologists. Here, the decision of radiologists first in line can be observed by radiologists second in line, which increases the identifiability and the evaluation potential of individual contributions made at the first checking position. In addition, the second-in-line radiologists can infer the reliability of preceding radiologists by reviewing their decisions. In case radiologists first in line work highly reliable, it is possible that the second-in-line radiologists adapt their behavior, based on this information, and invest less effort in the task. Consequently, all the processes mentioned here may trigger social loafing to a varying extent. It seems likely that blinded processes have a lower risk of social loafing compared to non-blinded processes. In the non-blinded process, the risk of social loafing might be most pronounced in the second checking position in case the first checker is perceived as reliable. If social loafing occurs in one or even both positions in sequential redundant work settings, this may diminish or even nullify the gains expected from utilizing redundancy principles. For example, if social loafing causes the individual reliability of two people to decline from 70% to 45% each, their combined reliability will reach only 69.8%, representing a lack of improvement over a single-person setting. In cases of less severe reductions of individual reliability, human redundancy might still provide advantages, but not at the expected high level.

Thus far, very little experimental research is available that has investigated how possible social loafing effects might affect performance in operational human-redundancy settings. Most of the research in this area has addressed the case of parallel redundancy, comparing two persons that are responsible for a given (monitoring) task simultaneously with persons being solely responsible for the task (Cymek, 2018; Domeinski, Wagner, Schoebel, & Manzey, 2007; Mosier, Skitka, Dunbar, &

McDonnell, 2001; Skitka, Mosier, Burdick, & Rosenblatt, 2000). The results suggest that as soon as a second person is added, parallel redundancy indeed might cause a reduction of individual effort and/or performance, which may even fully inhibit a higher overall reliability (Cymek, 2018; Mosier et al., 2001; Skitka et al., 2000). However, little is known about whether social loafing also occurs in *sequential redundancy*. A few anecdotal reports from the field (Bertovic, 2015) and a couple of interview studies (Armitage, 2008; Dickinson, McCall, Twomey, & James, 2010) do indeed hint at social loafing in sequential human redundancy. Nevertheless, despite being the prevalent form of redundancy in real-world settings, only a few experimental studies have thus far investigated the specific effects of sequential human redundancy on performance, with a somewhat inconsistent pattern of results. For example, one set of studies in health care addressed this issue with respect to double reading in mammography. Specifically, these studies investigated whether double reading (be it a non-blinded or blinded procedure) led to better detection performance than single reading (Denton & Field, 1997; Karau & Williams, 1993; Taplin et al., 2000; Thurfjell, 1994; Thurfjell, Taube, & Tabár, 1994); and one study even directly compared the effects of non-blinded vs. blinded procedures (Klompenhouwer et al., 2015). Altogether, these studies only found very small differences, if any, between double- and single-reading procedures, and only a marginal advantage of blinded double reading over non-blinded double reading. Based on these results, it is questionable whether a second radiologist should actually screen the X-rays, and whether the choice of a blinded or non-blinded process matters much. Unfortunately, these studies only reported the detection performance, and did not report any effort-related data, such as inspection time, search intensity, gaze behavior, or subjective effort ratings for the task. Thus, it is not absolutely certain whether it was reduced motivation that prevented a better team performance, or whether, for example, the non-perfect reliability of single radiologists was already the maximum achievable performance, as certain cancers were practically undetectable. Either way, these effects might also be taken as

indications of possible social loafing effects in these settings, as two sequentially working radiologists did not significantly outperform one radiologist.

Up to now, our literature search has only uncovered a single pilot study that explicitly investigated the effects of sequential human redundancy on effort and performance, which adopted a socio-psychological perspective (Conte & Jacobs, 1997). In this laboratory study, students had to perform a blinded error-checking task either alone or redundantly before or after one or two agents of varying reliability, and with a varying degree of accountability for the task. The study found that system characteristics, such as the presence or absence of redundancies and individual accountability, explained significant variance in measures of accuracy (7%) and time spent on the task (9%). Here, accuracy was a measure of performance, while the time invested served as a measure of effort. These results indicate that redundant systems can influence human effort and performance. Unfortunately, due to a large number of experimental factors and an incomplete design structure, no systematic assessment of how each specific factor (reliability of coworker, number of redundancies, accountability) affected effort and performance in the different checking position was possible. Likewise, Conte and Jacobs (1997) missed to investigate whether the joint team performance was actually better than that of individuals working in a non-redundant setting. However, such a comparison would be important to judge the practical consequence of adding individuals in an effort to increase overall reliability.

This study attempts to address performance consequences of sequential redundancy in a more systematic way. We conducted two different experiments which simulated a quality-control task that was performed either alone or by two participants in a sequentially redundant way. Redundant participants were informed that the quality control was performed twice and that an insufficient artifact was sorted out as soon as one of them found fault with it. Experiment 1 used a blinded procedure, where the work results were not shared between team partners and where no information about the counterpart's reliability was given. Here, social loafing could theoretically

occur due to the redundant task completion; however, we did not hypothesize any difference between the two positions, since they were both equally low-identifiable and unaware of each other's reliability. Experiment 2 used a non-blinded procedure, where the decisions of participants first in line were transmitted to their subsequent team partners. Here, the independence of both team partners was substantially reduced. The results of participants first in line were evaluable by participants second in line, which may hinder social loafing in the first position. Moreover, the second-in-line participants could derive reliability information concerning the participants preceding them by rechecking the decisions taken in the first position. If participants first in line performed highly reliably, this could lead to more social loafing in the second position, especially in the course of time as trust in the ability of the first-in-line participants grows. In both experiments, subjective and objective effort measures were taken, revealing how much checking effort each participant invested in the quality-control task. These measures are complemented by the failure-detection performance as a reduced checking behavior raises the risk for missing failures. In addition, objective effort, and failure-detection performance in the non-redundant condition were contrasted with the overall team effort and team-failure-detection performance in the redundant condition to get insights into the effects of redundancy.

EXPERIMENT 1: BLINDED QUALITY CONTROL

The first experiment investigated the effects of sequential human redundancy on subjects' effort and performance in a blinded quality-control task, where no information from the first team partner (Rpos1) was transferred to the second team partner in a sequence (Rpos2), and participants did not receive any explicit information concerning their team partners' reliability; they were only told whether they were working at the first or the second position. This experiment tested the general assumption that human sequential redundancy leads to social loafing at both positions in the sequence, compared to a condition where only one individual is responsible for the task. This assumption was based on the fact that the perceived correlation between individual input and

valued outcome is lower in group tasks than when a person conducts the task alone. We thus hypothesized that both the first-in-line and second-in-line participants might show signs of social loafing, which should be reflected in a reduced checking behavior (effort) compared to the non-redundant participants and may consequently even lead to a lower failure-detection performance (H1). However, effort and performance in both checking positions were not expected to differ from each other, since the factors that have been found to impact social loafing did not vary according to position in the sequence (H2). For example, this holds true for the individual performance being equally identifiable at both positions, and individuals working in either position having no knowledge or insight about how reliably their team partners worked. Furthermore, we did not expect social loafing effects as strong as that they would fully offset any redundancy effect. Thus, we hypothesized that the combined team effort and performance would still remain better than the effort and performance of a person being solely responsible for quality control (H3). This assumption was based on the rather small size of social loafing effects found in the laboratory study of Conte and Jacobs (1997) and in the meta-analysis of Karau and Williams (1993). In addition, to these hypotheses-based questions, two other aspects were addressed in an exploratory manner. The first one included the question to what extent a reduction of effort, indicated by less checks, would represent the result of a conscious decision. This is interesting as previous social loafing studies, addressing this question, found no consistent answer. We, therefore, also collected subjective assessments of effort invested in the task. Another aspect included possible changes of effort and performance across time. This aspect was addressed through the comparison of the different variables across different blocks of trials.

Methods

Transparency and Openness

Ethics. This study was approved by the local ethics committee at the Department

of Psychology, Technische Universität Berlin, Germany and is in accordance with the APA's Ethical Principles. Participants took part voluntarily, gave their informed consent, and were debriefed after the experiment.

Data. The data is available via the Open Science Framework (OSF) at https://osf.io/dga2n/?view_only=e783569673ef483f9d9521e83428a476.

Analytic Methods. The analytic code needed to reproduce analyses is available via the OSF at https://osf.io/dga2n/?view_only=e783569673ef483f9d9521e83428a476.

Materials. The materials are not available, but video records of the experimental environment are provided to support understanding.

The experimental environment

The multitasking program *MTOPS-R* (Multi-Task Operator Performance Simulation for Redundancy Research), consisting of three different tasks that need to be performed concurrently, was used for this experiment. One of these three tasks was a quality-control task that was conducted either in a redundant or non-redundant manner; the other tasks were diversions. All tasks were framed to simulate the demands that an operator of a chemical plant would have to deal with. The user interface is shown in Figure 1.

The first task, depicted in the upper left quadrant of the interface, was a resource-ordering task (ROT) in which participants had to ensure the supplies of chemicals to keep plant processes running. The resource-ordering task represented an easy mental arithmetic task. To order supplies, participants first had to subtract current stock ("Reserve"; e.g., 320 tons) from a required quantity ("Demand"; e.g., 500 tons), then type the difference into an order field, and finally submit their order. Each chemical was displayed for 15 seconds. Either after the submission of an order or after the time was up, the chemical disappeared and, after a delay of three seconds, a new chemical (abbreviated with labels such as K0-44BM, T-061MQ, etc.) arrived. If an order was not initiated, the

chemical reaction for which the substance was needed was canceled. This routine was continued until the experiment ended.

The second task, displayed in the upper right quadrant, was a coolant-exchange task (CET). Participants had to exchange coolant in a two-vessel cooling system as quickly as possible to prevent the overheating of the plant. For this purpose, the different valves (⋈) had to be opened and closed in a defined sequence to drain used coolant and to refill the vessels with fresh coolant. Participants could tell by the color of the vessel whether it contained no coolant (gray), used coolant (green), or fresh (blue) coolant. The filling and emptying speed differed from exchange to exchange, and only one inlet and one outlet valve could be opened at the same time. The minimum time for a complete exchange cycle of both vessels ranged from 14 to 32 seconds.

The third task was displayed in the lower right quadrant and comprised the quality-control task (QCT), which was most relevant for this study. Here, participants were required to inspect the pressure and the pH value of a fixed number of containers before dispatch. Both parameters could be in a safe range, too high, or too low. To check the pressure, participants needed to open an image showing a certain number of dark red spots on a light red background (Figure 2, upper left and right). Ten spots indicated that the pressure in a container was acceptable, and eleven or nine spots indicated unacceptably high or low pressure, respectively. To check the pH value, participants had to select the “pH-value” tab and then compare a color-coded pH value with a reference scale (Figure 2, lower right). A pH value between six and ten was said to be safe, whereas a pH value below six was too low, and a pH value above ten was too high. It is important to note that it took some effort to access all parameter displays, but that the visual assessment of both parameters was clearly answerable, as the pictures were not blurred or did not contain any distracting noise. Thus, all human beings who see colors should be able to solve the task without setting off false alarms. If participants found a parameter that was unacceptable, they had to change the default setting (everything is okay) from a dropdown menu accordingly (Figure 2, lower left). If no checks or menu

changes were made within nine seconds, the current container disappeared, and the next container appeared after a random time period of two to ten seconds. If participants inspected a container, additional time was warranted for each parameter check to ensure that after inspection, a subsequent change in the dropdown menu would be feasible. The maximum display time for a container was set to 30 seconds. According to this logic, conscientiously checking the containers prolonged the total working time for the task.

Depending on the experimental conditions, the interface and procedure of the quality-control task varied slightly. In the redundant condition, a network symbol was added to the interface next to the dropdown menu, which flashed green when data between the first checker and the second checker was supposedly transmitted (Figure 1, lower right). The first checkers therefore saw a blinking green network symbol when a container left their desk, whereas the second checkers saw this signal just before a container arrived at their desk.

Design

A 3 (condition) x 3 (block) mixed-factorial design with repeated measurements of the second factor was used for the experiment. The first factor included three different conditions. One-third of the participants worked alone on all three MTOPS-R tasks (NonR), and the remaining two-thirds of participants worked alone on the resource-ordering and coolant-exchange tasks, but in a sequential redundancy setting in the quality-control task, either as the first (Rpos1) or as the second team member in line (Rpos2). The second factor represented a within-subjects factor that was included to exploratively investigate whether checking effort and/or possible social loafing effects were influenced by time-on-task. Participants were presented a minimum of 90 containers in the quality-control task. For the data analysis, this number was split into three blocks of 30 containers each. However, this block structure was not transparent to the participants.

Procedure

Four participants, assigned to two workstations, performed the experiment simultaneously. Each workstation included two PCs that were separated by partition walls so that the participants could not see each other. Upon arrival, participants were randomly assigned to the different cubicles, they gave their informed consent, filled out a demographic survey, and received written instructions explaining the specific experimental condition and the different tasks involved. Participants of the non-redundant group were instructed that they would work on all three tasks alone. In the two redundant conditions, participants were instructed that they were solely responsible for the coolant-exchange and resource-ordering task, but that they would perform the quality-control task redundantly with a second participant for reasons of safety (i.e., to increase the probability of detection of unacceptable containers). They were also told that the results of the participants first in line would not be forwarded to the participants second in line in order to increase the level of independence in the quality-control task. Furthermore, they were instructed that only their integrated team performance was decisive, i.e., that all containers would be excluded before dispatch that had been marked unacceptable by either or both of them. The participants were additionally informed that the workstations were interconnected, to support the teamwork in the quality-control task, and what their position would be in the sequential procedure.

After reading the illustrated and written instructions, the participants first practiced the three tasks separately to learn how they had to be performed. Thereafter, the participants completed a paper questionnaire to check their understanding. After clarifying any misunderstandings or remaining questions, the data collection started.

While the participants in the non-redundant condition immediately started performing the tasks, the participants in the two redundant conditions initially needed to connect their computers by logging in with a given IP address provided on their monitor. As soon as both team partners had logged in, the program started for both of them simultaneously. It is important to note that, in reality, the computers were not connected. To further reinforce the belief that both team partners

were really working sequentially on the quality-control task, the first checker worked on all three tasks from the very start, whereas the second checker was given the resource-ordering and coolant-exchange task right away but received the first container only after a delay of two minutes (the same routine was used in the training session). Each participant first received an identical set of n = 90 containers. However, the checkers second in line received ten additional containers to ensure that even if a participant was putting little effort into the quality-control task, he or she would not finish the work on the multitask before the participant first in line. However, only the identical set of 90 containers was included in the data analyses, divided into three blocks of 30 containers each to define the block factor. Altogether, the teamwork manipulation seemed successful as there were no signs that participants did not believe that they were working in teams of two.

A total of 18 out of the 90 containers (20%) presented to the participants were "unacceptable," i.e., had either a pH value or a pressure that was not in the safe range. A description of the distribution of unacceptable containers during the experiment is provided in Table 1.

In total, each experimental session took just under 40 minutes (M = 37.1 SD = 2.6). At the end of each session, participants' multitasking preference (Multitasking Preference Inventory of Poposki and Oswald (2010)) was assessed as a control variable, and subjective variables were taken in order to compare subjective and objective measures of effort. Participants were debriefed and received their compensation.

Dependent variables

Three sets of dependent variables were reviewed. The first set of variables included individual effort and/or performance measures for the three tasks and corresponded to the variables used in previous redundancy research using MTOPS-R (Cymek, 2018; Domeinski et al., 2007). These were derived from log files in which all actions performed by a participant during the experiment were recorded (mouse clicks, keystrokes, etc.). Performance in the resource-ordering task was measured by the total number of resource orders per block. The performance of the coolant-exchange task was

assessed by the number of coolant-exchange cycles performed per block. In the latter two tasks, the selected performance measures were strongly related to the effort invested in the task as, for example, the opening and closing of the valves in the coolant-exchange task directly determined the number of coolant-exchanges per block. However, in the quality-control task, effort and performance were less directly linked, so that a distinction was made between measures of effort and performance. The effort measure was defined as the number of "safe container checks" and included all intact containers of each block for which a participant checked both relevant parameters, as well as all unacceptable containers for which at least the relevant/deviating parameter was checked. As a performance measure the failure-detection performance served, which was defined as the number of container deviations identified out of the total of 18 unacceptable containers.

The second set of variables represented effort and performance measures in the quality-control task on the team level. Here, effort and performance within pairs were aggregated for each of the two individual quality-control variables. A "safe container check" at the team level was defined as any instance where both relevant parameters (for intact containers) or at least the relevant/deviating parameter (for unacceptable containers) were checked by at least one of the two team members. The failure-detection performance on the team level was defined as the number of unacceptable containers that were identified by at least one team member.

Finally, the third set of dependent variables comprised a set of three subjective effort ratings in the quality-control task. Participants had to indicate on a seven-point rating scale how much they agreed to different effort statements to learn whether possible objective effort reductions corresponded with the perceptions of the participants. Beside the rather general item "I made an effort in the quality-control task." (#SE1), which all participants had to rate in order to check whether they set a comparable priority on this task within the multitasking scenario, two more specific items were presented only to participants that worked redundantly. One item asked to judge whether the same amount of effort would have been invested in a non-redundant setting (#SE2: "I made as much

effort in the QCT as I would have if I had been solely responsible for the task."), and the other directly asked whether the redundant setting led to more reliance and less subjective checking in order to reduce workload (#SE3: " Sometimes I relied on the performance of my team partner to have a rest.").

Participants

A total of *n* = 60 participants recruited out of a university participant pool took part in the experiment and were randomly assigned to the three conditions. Both students and non-students are registered in the participants' pool, and most of them frequently participate in research to receive monetary compensation. Based on a G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) calculation, the chosen sample size should suffice to detect large between-subject effects and medium within-subjects and interaction effects in our ANOVAs (α err prob = 0.05, 1-β err prob = .95). However, two participants (one of the NonR and one of the Rpos1 condition) had to be excluded from the data analyses since they largely disregarded the quality-control task, which led to extreme deviations of their performance from the mean of their specific group. Thus, the final sample included in the data analyses consisted of 58 participants. Of those 58 participants, 30 were female; 44 were students (all without psychology backgrounds). Their ages ranged from 20 to 35 (*M* = 26.0, *SE* = 0.6). Participants in all experimental conditions did not differ regarding their preference for multitasking, $F(2, 55) = 0.03$, $p = .973$, $\eta_p^2 < .01$. Participants were compensated with €18.

Results

The resource-ordering task

The means and standard errors of performance scores for the resource-ordering task (ROT) are shown in Figure 3. A 3(group) x 3(block) mixed ANOVA revealed a significant main effect of block, $F(2, 110) = 35.27$, $p < .001$, $\eta_p^2 = .39$, as slightly more resources were ordered with time. Furthermore, a significant interaction of group and block emerged, $F(4, 110) = 6.38$, $p < .001$, $\eta_p^2 = .19$, as participants of the Rpos2 group did not improve their performance over time as much as

the NonR and Rpos1 participants. However, no main effect of group was found, $F(2, 55) = 0.82$, $p = .444$, $\eta_p^2 = .03$.

The coolant-exchange task

The means and standard errors of performance scores for the coolant-exchange task (CET) for the three groups in the different blocks are shown in Figure 4. A 3(group) x 3(block) mixed ANOVA revealed a significant main effect of block as more coolant-exchange cycles were made with time, $F(1.71, 93.92) = 15.95$, $p < .001$, $\eta_p^2 = .23$. Neither a main effect of group, $F(2, 55) = 0.469$, $p = .628$, $\eta_p^2 = .02$, nor an interaction effect was found, $F(3.42, 93.92) = 1.583$, $p = .215$, $\eta_p^2 = .05$.

The quality-control task: individual level

The mean numbers of safe-container checks for the different groups and the three blocks of the experiment are shown in Figure 5 (left). Participants in all conditions checked containers with a similar intensity (NonR: $M = 22.3$, $SE = 1.3$, 74%; Rpos1: $M = 25.1$, $SE = 1.0$, 84%; Rpos2: $M = 23.8$, $SE = 1.2$, 79%), and this effort level did not change much across the three blocks. According to the 3(group) x 3(block) mixed-effect ANOVA, neither the main effect of group, $F(2, 55) = 1.81$, $p = .174$, $\eta_p^2 = .06$, nor the main effect of block, $F(1.51, 83.05) = 1.90$, $p = .165$, $\eta_p^2 = .03$, nor the group x block interaction, $F(3.02, 83.05) = 0.81$, $p = .491$, $\eta_p^2 = .03$, became significant. Similarly, only small differences among the three groups emerged for the failure-detection performance. A Kruskal-Wallis test comparing the failure-detection performance of the NonR ($M = 13.1$, $SE = 0.7$, 73%), Rpos1 ($M = 14.8$, $SE = 0.8$, 82%), and Rpos2 group ($M = 13.9$, $SE = 0.6$, 77%), just failed the conventional level of significance, ($H(2) = 5.52$, $p = .063$). Not even the general trend of differences between means with slightly higher detection rates under the redundant conditions than under the NonR conditions was in the expected direction.

The quality-control task: team level

The mean numbers of “safe container checks” at the team level and, for comparison, that of the NonR condition are shown for the three blocks in Figure 5 (right). A 2(group) x 3(block) mixed ANOVA

revealed a mean effect of group, indicating that the checking in the Rteam condition (M = 29.2, SE = 0.3, 97%) was significantly better than in the NonR condition (M = 22.3, SE = 1.3, 74%), $F(1, 36) = 32.96$, $p < .001$, $\eta_p^2 = .48$. The main effect of block was significant as well as checking intensity slightly improved with time, $F(1.32, 47.45) = 4.62$, $p = .027$, $\eta_p^2 = .11$, but no interaction was found, $F(1.32, 47.45) = 2.24$, $p = .134$, $\eta_p^2 = .06$. The better checking on the team level also led to a correspondingly higher failure-detection rate on the team level compared to the NonR condition. A Mann-Whitney-U test found that teams detected more failures (M = 17.2, SE = 0.3, 96%) than individuals from the non-redundant group (M = 13.1, SE = 0.7, 73%), $U = 19.5$, $p < .001$, $r = 0.77$.

The quality-control task: subjective ratings

The mean ratings of subjective effort are given in Table 2. One-way analyses of variance (ANOVA) found no differences between the experimental groups for all three items measuring subjective effort (all $F < 1.3$ and all $\eta_p^2 < .05$).

Discussion

Experiment 1 tested whether social loafing occurred in a quality-control setting that used blinded sequential human redundancy, and whether these effects jeopardized the anticipated gains in overall reliability. We hypothesized that social loafing could occur in the redundant conditions (H1), with an equally pronounced strength on both checking positions (H2). However, neither the objective effort and performance measures (i.e., number of safe container checks and failures detected, respectively), nor the subjective effort measures differed among the two redundant groups and the non-redundant group. Also, no notable changes of checking intensity across blocks indicating possible time-on-task effects were detected. Based on Conte and Jacobs (1997), who used a blinded process as well, we expected that at least some social loafing might occur. But because Conte and Jacobs (1997) attempted to account for many variables such as accountability for the task as well as the number, nature, and perceived reliability of coworking agents, a clear-cut comparison to our findings is not feasible. Since no performance differences emerged from the quality-control task,

differences between the groups in the concurrent tasks can be ignored, as they obviously do not indicate a rational effort shift towards the tasks of sole responsibility at the expense of effort invested in quality control. This suggests that sequential human redundancy does not necessarily lead to social loafing effects if it is implemented in a similar manner as it was in Experiment 1, where the relevant quality-control task was accomplished using a blinded procedure and where unfamiliar people worked with each other in a way that did not generate any information regarding the other's performance level.

Since no support of social loafing effects was found in our data, it came as no surprise that the mean team performance was superior to the performance of the non-redundant condition (H3). Almost all containers were safely checked by the teams (97% vs. 74% in NonR group) and almost all failures were detected (96% vs. 73% in NonR group). Thus, a clear benefit of the implementation of human redundancy emerged. Overall, the results of the first experiment suggest that blinded sequential human redundancy might help to increase the reliability of human performance in tasks such as quality control, which demand thorough visual inspection. However, many implementations of sequential human redundancy in practice do not follow a blinded process but include an information transfer between the different positions in the sequence, i.e., a transfer of the decision made by the first position to the second one. Such information transfer has two important implications: First, the performance of persons in the first position becomes identifiable by persons in the second position. Second, given this, persons in the second position have the possibility to infer the reliability of persons in the first position by reviewing their decisions. The second experiment addressed how this affected effort and performance in the two positions as well as the overall redundancy effect.

EXPERIMENT 2: NON-BLINDED QUALITY CONTROL

This experiment investigated the effects of sequential human redundancy on subject effort and performance in a non-blinded quality-control task in which decisions made by the first team member

were forwarded to the second team member. However, in order to keep the reliability information along with this information transmission constant, the second team members received not the actual decisions made at the first checking positions, but instead received a-priori-scripted checking decisions of a highly reliable first checker (96%). This was done to generate a situation that should increase the chances for participants second in line to start loafing. Our first hypothesis was that first-in-line participants would not engage in social loafing, as their individual contribution could be assessed by second-in-line participants (H1). However, second checkers, realizing that their preceding team partner already worked highly reliably, might become prone to social loafing effects over the course of time. This should be reflected in a reduced checking intensity, eventually even allowing more unacceptable containers to pass (H2). However, since substantial social loafing effects were only hypothesized to arise in the second position, we still assumed that the overall team performance would be somewhat better than the performance of a single checker, thus reflecting a perhaps diminished but nevertheless visible advantage of the redundant work setting (H3).

Method

Transparency and Openness

Ethics. This study was approved by the local ethics committee at the Department of Psychology, Technische Universität Berlin, Germany and is in accordance with the APA's Ethical Principles. Participants took part voluntarily, gave their informed consent, and were debriefed after the experiment.

Data. The data is available via the Open Science Framework (OSF) at https://osf.io/dga2n/?view_only=e783569673ef483f9d9521e83428a476.

Analytic Methods. The analytic code needed to reproduce analyses is available via the OSF at https://osf.io/dga2n/?view_only=e783569673ef483f9d9521e83428a476.

Materials. The materials are not available, but video records of the experimental environment are provided to support understanding.

Experimental environment

The MTOPS-R program was slightly adapted to test the non-blinded redundancy procedure. The difference was that the participants second in line saw their putative teammates' quality-control checking decision, indicated on a small display (Figure 6, lower left) that turned either green when the participants first in line reported no deviations, or red when the first checker reported that at least one of the parameters diverged from the safe range. It is important to note that these decisions did not correspond to the real-time decisions of the test subjects working first), but instead were based on a simulated script that was used to control for the reliability of Rpos1 decisions across participants. Consequently, the dropdown menu in the second checking position did not always display the default settings (both parameters in the safe range) but the supposed dropdown-menu selection of a scripted highly reliable first checker. Thus, the second participants in line only had to change the dropdown menu if additional container deviations were detected, or if they disagreed with the assessment results of the scripted first checker. Exemplary videos showing the interactions with MTOPS-R in the three conditions are available via the Open Science Framework at https://osf.io/dga2n/?view_only=e783569673ef483f9d9521e83428a476 (Cymek, 2021).

Design

The same 3 (condition) x 3 (block) mixed-factorial design was used as in the first experiment.

Procedure

The procedure followed that of Experiment 1. The main difference from Experiment 1 was that the results of a scripted highly reliable first checker were forwarded to participants second in line.

A total of 18 out of the 90 containers (20%) presented to the participants first in line were unacceptable, i.e., they had either a pH value or a pressure that was not in the safe range. Theoretically, the participants first in line could therefore each identify a maximum of 18 unacceptable containers. The second-in-line participants saw the same 90 containers (plus 10 additional containers). The manipulated indications they got from the first checking position included

a total of 95.6% correct assessments, i.e., green signals for the *n* = 72 containers that actually were intact and red signals for *n* = 14 containers that were unacceptable. However, *n* = 4 unacceptable containers were "missed" by the scripted highly reliable first checker, i.e., they were accompanied by a green signal transferred to Rpos2. This corresponded to a 77% detection rate for unacceptable containers of the scripted first checker, which simulated a sort of miss-prone checking performance in this position. A detailed description of the distribution of unacceptable containers over the experiment in the two different positions is provided in Table 3. Unacceptable containers missed by the scripted first checker only occurred in the second half of the experiment to mimic situations where a generally highly reliable coworker becomes temporarily tired or distracted.

Dependent variables

The same dependent measures as in Experiment 1 were used. However, this time the failure-detection performance was defined as the number of container deviations identified out of the unacceptable containers that appeared as unmarked in all groups (containers 65, 69, 82, and 90) to avoid biasing the failure-detection comparison. These four (out of 18) unacceptable containers theoretically could have been missed by Rpos2 participants as they received "green" signals for these containers. The 14 unacceptable containers, that were already marked with a "red" signal from the scripted first checker, were not included in the failure-detection rate, as the visual inspection was generally manageable, and the second-in-line team members predictably never reclassified these containers as falsely intact.

Participants

A total of *n* = 56 participants who were recruited from the same participant pool used in Experiment 1 took part in the experiment and were randomly assigned to the three conditions. However, four participants (three of the NonR and one of the Rpos1 group) had to be excluded from data analyses since they largely disregarded the quality-control task, which led to extreme deviations of their performance from the mean of their specific group. Thus, the final sample included in the

data analyses consisted of 52 participants, a sample size that should, according to a G*Power calculation (Faul et al., 2007), suffice to detect large between-subject effects and medium within-subject and interaction effects in our ANOVAs (α err prob = 0.05, 1-β err prob = .95). Of those 52 participants, 30 were female; 44 were students (without psychology backgrounds). Their ages ranged from 20 to 35 (*M* = 27.4, *SE* = 0.5). In all experimental conditions, participants did not differ regarding their preference for multitasking, $F(2, 49) = 0.50$, $p = .609$, $\eta_p^2 = .02$. Participants were compensated with €15.

Results

The resource-ordering task

The means and standard errors of performance scores in the resource-ordering task (ROT) for the three groups in the different blocks are shown in Figure 7. A 3(group) x 3(block) mixed ANOVA revealed a significant main effect of block as the number of resources ordered per block increased across blocks, $F(1.46, 71.48) = 49.94$, $p < .001$, $\eta_p^2 = .51$. No main effect of group, $F(2, 49) = 0.33$, $p = .724$, $\eta_p^2 = .01$, and no interaction effect was found, $F(2.92, 71.48) = 1.23$, $p = .304$, $\eta_p^2 = .05$.

The coolant-exchange task

The means and standard errors of performance scores for the coolant-exchange task (CET) for the three groups in the different blocks are shown in Figure 8. A 3(group) x 3(block) mixed ANOVA revealed a significant main effect of block, $F(1.62, 79.42) = 57.45$, $p < .001$, $\eta_p^2 = .54$, as all groups performed more coolant exchanges with time. No main effect of group, $F(2, 49) = 0.50$, $p = .609$, $\eta_p^2 = .02$, and no interaction effect was found, $F(3.24, 79.42) = 2.11$, $p = .101$, $\eta_p^2 = .08$, for this task either.

The quality-control task: individual level

The mean number of “safe container checks” performed by participants in the different conditions are shown in Figure 9 (left). A 3 (group) x 3(block) mixed ANOVA revealed a significant

main effect of group, $F(2, 49) = 10.15$, $p < .001$, $\eta_p^2 = .29$. Planned contrasts revealed that participants in the Rpos2 group performed significantly fewer safe container checks (M =18.0, SE = 2.1, 60%) than NonR participants (M = 24.3, SE = 0.8, 81%), Bonferroni-Holm adjusted p = .004. In contrast, no difference was found between the mean checking behavior of Rpos1 (M = 26.0, SE = 0.8, 87%) and NonR participants, p = .380. A significant main effect for block, $F(1.65, 80.64) = 4.35$, p = .022, partial η^2 = .08, emerged as well. While participants working alone or at the first position in the redundant setting slightly improved their checking intensity over time, the opposite occurred for participants at the second position in the redundancy setting. These opposing trends were confirmed by the significant interaction between group and block, $F(3.29, 80.64) = 3.67$, p = .013, partial η^2 = .13.

The detected difference in the checking behavior also led to different levels of failure-detection performance. Assessment of detection performance was only based on the four containers that theoretically could have been missed by Rpos2, i.e., those unacceptable containers for which participants in this position received simulated erroneous “green” signals from the first checking position. The detection performance for these containers was lowest in the Rpos2 condition (M = 2.5, SE = 0.4, 63%), highest in the Rpos1 (M = 3.7, SE = 0.2, 93%), and in between in the NonR condition (M = 3.2, SE = 0.2, 80%). A Kruskal-Wallis test comparing the three conditions revealed that there was a significant difference among these groups, ($H(2) = 6.32$, p = .043). However, the effect was relatively weak and pairwise comparisons found that only the detection performance in the Rpos1 condition was significantly higher than in the Rpos2 condition (Bonferroni-Holm adjusted p = .036, r = 0.43), but neither the Rpos1 (p = .136, r = 0.26) nor the Rpos2 group (p = .322, r = 0.16) differed significantly from the NonR group in this respect.

The quality-control task: team level

The mean numbers of “safe container checks” on the team level and, for comparison, that of the NonR condition are shown for the three blocks in Figure 9 (right). As becomes evident, the checking intensity of Rteam (M = 27.9, SE = 0.6, 93%) and NonR participants (M = 24.3, SE = 0.8, 81%) differed

significantly, $F(1, 32) = 15.43$, $p < .001$, $\eta_p^2 = .33$. A main effect of block, $F(2, 64) = 4.70$, $p = .012$, $\eta_p^2 = .13$, was found as well, reflecting a slight increase of checking effort across the three blocks. However, no interaction effect was found, $F(2, 64) = 1.89$, $p = .161$, $\eta_p^2 = .06$. A similar difference between Rteam and NonR was also found for failure-detection performance, which was perfect on the team level ($M = 4.0$, $SE = 0.0$, 100%) and, thus, considerably better than the 80% detection rate in the NonR group, $U = 221.00$, $p < .001$, $r = 0.58$.

The quality-control task: subjective ratings

Mean ratings of subjective effort are given in Table 4. One-way analyses of variance (ANOVA) found no differences between experimental conditions for item #SE1 (“I made an effort in the quality-control task.”), $F(2, 49) = 1.63$, $p = .207$, $\eta_p^2 = .06$, and item #SE2 ("I made as much effort in the quality-control task as I would have, if I had been solely responsible for the task.”), $F(1, 33) = 4.07$, $p = .052$, $\eta_p^2 = .11$. However, item #SE2 was very close to reaching significance. Only for item #SE3 (“Sometimes I relied on the performance of my team partner to have a rest.”) was a significant effect found, as Rpos2 participants agreed more than Rpos1 participants to this item, $F(1, 33) = 5.90$, $p = .021$, $\eta_p^2 = .15$.

Discussion

The results of Experiment 2 show that human sequential redundancy may indeed be prone to social loafing if information is transferred between two redundantly working individuals. In our data, evidence for such effects was especially found in the condition where participants served as second checkers in the quality-control task. Participants working in this position did not check as many containers as participants working alone; this difference even increased across time. In contrast, participants working in the first position of the redundant team did not differ from participants working alone. On a descriptive level, the former participants even improved their performance, i.e., checked more containers than the participants in the non-redundant condition. It is important to note that the decreased checking performance in the second checking position was not accompanied

by increased performance in the concurrent tasks and, thus, does not seem to simply be a trade-off between tasks, but a real loafing effect in terms of reduced quality-control effort. This latter effect had direct consequences for the detection of unacceptable containers. Out of the four containers that had not already been identified by the first simulated checker, on average only 2.5 (63%) containers were identified by the participants second in line. In comparison, the detection rates for these containers were considerably higher in both the NonR (80%) and the Rpos1 condition (93%). The very high detection rates in the first position seem to be directly related to the relatively high effort exerted in checking at this position, even compared to the non-redundant condition. All of these findings are generally in line with our predictions derived from social loafing research. To reiterate, we expected that less or essentially no social loafing would occur in the first checking position because, due to the non-blinded procedure, this person's checking decisions would be identifiable and could be evaluated by the second person in line (H1). In the second checking position, however, we expected social loafing since that person's contribution to the group outcome was less identifiable, and the motivation to put major effort into the task was increasingly diminished by the ongoing experience that a very reliable first checker had already checked the containers (H2).

Interestingly, we found that the objective effort performance was in line with the subjective effort ratings. Participants from all conditions agreed that they generally made an effort in the quality-control task. This suggests that the participants did not differ much in their priority setting when working concurrently on the three tasks. Such an interpretation is also in line with the performances in the two other tasks (resource ordering, coolant exchange) which did not differ significantly either between the three groups. Yet, when asked more specifically about social loafing over the course of the experiment ("Sometimes I relied on the performance of my team partner to have a rest"), participants working in the second position agreed with this statement more than participants working in the first checking position. This fits to the finding of less checks performed by the participants in this condition and suggests that the objective effort reductions found in the

number of performed checks were not the result of a subconscious behavioral adaptation, but the results of a deliberate decision. Further support for such interpretation of social loafing in our experiment as a conscious process is provided by the lower ratings on the item "I made as much effort in the quality-control task as I would have, if I had been solely responsible for the task." which only just missed the conventional level of statistical significance.

In order to investigate the extent to which the social loafing effect observed in the second checking position was strong enough to compromise the overall reliability effect of redundancy, individual checking and failure-detection rates in the non-redundant group were compared with the combined checking and detection rate on the team level in the redundant condition. Even with the social loafing effect present, the mean team-checking performance still was significantly superior to that of the non-redundant group (NonR: 81%; Rteam: 93%). On top of that, the failure-detection performance was perfect on the team level, while participants working in the non-redundant condition only identified about 80% of the failures on average. Thus, even though a social loafing effect occurred in the second checking position, the overall team performance was not compromised and was still clearly superior to the performance of the non-redundant group, with a perfect detection of all unacceptable containers on the team level. These findings support Hypothesis 3. However, it should be noted that this effect obviously benefited from the increase in checking effort from the participants first in line which, at least to some extent, compensated for the social loafing effect in the second position.

GENERAL DISCUSSION

Sequential human redundancy is present in a variety of work environments and is often implemented to increase the accuracy of decisions that may have serious consequences. Unfortunately, social mechanisms exist that could limit this aspiration. Specifically, it has been an open question whether people who work according to a double-check procedure adapt to the team setting by reducing their individual effort, and whether these effort reductions can jeopardize the

anticipated benefits of the double-check procedure. This study found social loafing effects only when the implementation of sequential redundancy followed a non-blinded procedure, and when predominantly correct decisions made in the first position were transparent to the second position. In this case, participants working in the second checking position were especially likely to reduce their individual checking effort. However, even with the non-blinded procedure, this social loafing effect did not lead to an invalidation of the expected redundancy gain, but just a slight diminishing of what would have been an ideal redundancy effect. In contrast, in Experiment 1, where sequential redundancy in a blinded quality-control task was used, no social loafing in either position of the sequence was found at all, which led to an almost perfect checking and detection performance on the team level. This suggests that even when human beings work redundantly and experience a lower or unknown correlation between own personal input and group outcome, it does not automatically cause social loafing in a safety-relevant task. It seems that increasing the independence of components, by reducing information transmission between team members as much as possible, may even encourage subjects to work as if they were alone on the task.

Overall, this pattern of results found in the two experiments for the individual effort and performance consequences of sequential redundancy provide some support for the validity of predictions from the *Collective Effort Model* (CEM, Karau and Williams, 1993), although these experiments were not specifically planned as a validation of this model. In particular, our experiments support the prediction that the risks of social loafing effects in redundant work settings should be especially high if employees do not see a clear correlation between their own effort and the overall team performance. This was especially the case for the second checkers who worked in the non-blinded process after a highly reliable checker at the first position.

From an applied perspective, our findings suggest that work settings based on sequential redundancy can indeed induce social loafing effects. However, the strength of these effects found in our experiments were not strong enough to fully offset any benefits gained from redundancy. This

marks a contrast to previous studies that investigated possible social loafing effects induced by parallel redundancy. In these studies, the effects pointing to reduced individual effort induced by redundancy were sometimes so strong that the overall team performance was no better than that of single individuals (Cymek, 2018; Mosier et al., 2001; Skitka et al., 2000). One possible reason for this difference might be that the task setting of sequential redundancy automatically makes individuals feel more responsible for accomplishing their task than in parallel redundancy, as each individual works on the task timely independent of the other and hast to explicitly provide an own judgement/decision in order to proceed. Thus, the trigger to distribute responsibility and to engage in social loafing might be considerably reduced in sequential compared to parallel human redundancy. Such explanation would be in line with past research on parallel redundancy that found that specific treatments inducing a higher level of accountability can help to reduce social loafing effects (Mosier et al. 2001). Based on this explanation, it would be worth considering converting parallel redundant operations into sequential redundant operations, in cases where this is theoretically and practically possible. For example, it might be feasible that each pilot monitoring the autopilot is frequently asked to judge the current state of the system to increase each individual's accountability, rather than responding only to very rare alarms and system malfunctions.

Having described these possible applied consequences of our research, we must admit that one limitation of the experiments described here relates to the fact that they were conducted based on a low-fidelity simulation in the laboratory. Thus, the generalizability to real operational tasks conducted by experienced operators might be challenged. However, the fact that a lack of redundancy effect has also been reported in anecdotal reports from the field (Armitage, 2008; Bertovic, 2015) suggests that these effects are not limited to the laboratory. While such criticism might not fully be ruled out, we would at least perceive our results as providing some initial hints about possible risks of social loafing in sequentially redundant work settings, depending on the kind

of implementation (blinded – non-blinded) and position of the checker (first vs. second). Clearly, more applied research would be needed to externally validate these findings in the field.

From a more fundamental perspective, several follow-up questions arise that might also be addressed in future research. One research direction could be a direct comparison of sequential and parallel redundancy. This could test whether a task that is realized with sequential human redundancy leads to a greater sense of responsibility and more invested effort than when parallel redundancy is used for the same task. The authors believe that this might be possible as in sequential human redundancy each individual works on the task timely independent and has to provide a formally independent decision to proceed.

In addition, capitalizing on the current research which has included highly reliable first checkers, it might be interesting to investigate how an only moderately reliable first checker might impact the effort and performance of second checkers in a non-blinded process. Theoretically, in the latter case, the individuals second in line might perceive that their input has a higher impact on the group product, than when they follow a highly reliably team partner. The question here would be, whether social loafing occurred in such a setting at all and, if so, what effect this would have on team performance and the envisioned gain in reliability. This research would allow us to gain a more complete picture of social loafing as a function of different reliability levels of coworkers.

Future research could and should also test blinded and non-blinded double checks on stimuli that are noisy and more difficult to judge, as this the case when, for example, radiologists scrutinize mammograms. Furthermore, it would also be important to investigate teams with members who know each other, as this is a common feature in the field. The predictions concerning social loafing in such contexts would certainly be much more complex. Thus, the current research should only be considered as a first step towards investigating the effects of sequential human redundancy, which will hopefully stimulate continuing research.

Authors' Contribution

DHC designed the study in coordination with DM. DHC collected and analyzed the data and wrote the initial draft of the manuscript. DHC and DM critically revised the manuscript.

Acknowledgements

We would like to thank Marcus Bleil for technical support, and all reviewers and editors for helpful comments.

References

Allport, F. H. (1920). The influence of the group upon association and thought. *Journal of Experimental Psychology*, *3*(3), 159–182. https://doi.org/10.1037/h0067891

Anttinen, I., Pamilo, M., Soiva, M., & Roiha, M. (1993). Double reading of mammography screening films-one radiologist or two? *Clinical Radiology*, *48*(6), 414–421. https://doi.org/10.1016/S0009-9260(05)81111-0

Armitage, G. (2008). Double checking medicines: defence against error or contributory factor? *Journal of Evaluation in Clinical Practice*, *14*(4), 513–519. https://doi.org/10.1111/j.1365-2753.2007.00907.x

Bertovic, M. (2015). *Human Factors in Non-Destructive Testing (NDT): Risks and Challenges of Mechanised NDT* (Dissertation). Technische Universität Berlin, Berlin. Retrieved from https://depositonce.tu-berlin.de/bitstream/11303/4982/1/bertovic_marija.pdf

Clarke, D. M. (2005). Human redundancy in complex, hazardous systems: A theoretical framework. *Safety Science*, *43*(9), 655–677. https://doi.org/10.1016/j.ssci.2005.05.003

Conte, J. M., & Jacobs, R. R. (1997). Redundant Systems Influences on Performance. *Human Performance*, *10*(4), 361–380. https://doi.org/10.1207/s15327043hup1004_3

Cymek, D. H. (2018). Redundant Automation Monitoring: Four Eyes Don't See More Than Two, if Everyone Turns a Blind Eye. *Human Factors*, 1-20. https://doi.org/10.1177/0018720818781192

Cymek, D. H. (2021). Sequential Human Redundancy. Retrieved from https://osf.io/dga2n/?view_only=e783569673ef483f9d9521e83428a476.

Denton, E., & Field, S. (1997). Just how valuable is double reporting in screening mammography? *Clinical Radiology*, *52*(6), 466–468. https://doi.org/10.1016/S0009-9260(97)80010-4

Dickinson, A., McCall, E., Twomey, B., & James, N. (2010). Paediatric nurses' understanding of the process and procedure of double-checking medications. *Journal of Clinical Nursing*, *19*(5-6), 728–735. https://doi.org/10.1111/j.1365-2702.2009.03130.x

Domeinski, J., Wagner, R., Schoebel, M., & Manzey, D. (2007). Human Redundancy in Automation Monitoring: Effects of Social Loafing and Social Compensation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *51*(10), 587–591. https://doi.org/10.1177/154193120705101004

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Harkins, S. G. (1987). Social loafing and social facilitation. *Journal of Experimental Social Psychology*, *23*(1), 1–18. https://doi.org/10.1016/0022-1031(87)90022-9

Harkins, S. G., & Jackson, J. M. (1985). The Role of Evaluation in Eliminating Social Loafing. *Personality and Social Psychology Bulletin*, *11*(4), 457–465. https://doi.org/10.1177/0146167285114011

Harkins, S. G., & Szymanski, K. (1989). Social loafing and group evaluation. *Journal of Personality and Social Psychology*, *56*(6), 934–941. https://doi.org/10.1037//0022-3514.56.6.934

Karau, S. J., & Williams, K. D. (1993). Social loafing: A meta-analytic review and theoretical integration. *Journal of Personality and Social Psychology*, *65*(4), 681–706. https://doi.org/10.1037/0022-3514.65.4.681

Kerr, N. L., & Bruun, S. E. (1983). Dispensability of member effort and group motivation losses: Free-rider effects. *Journal of Personality and Social Psychology*, *44*(1), 78–94. https://doi.org/10.1037/0022-3514.44.1.78

Klompenhouwer, E. G., Voogd, A. C., den Heeten, G. J., Strobbe, L. J. A., Haan, A. F. J. de, Wauters, C. A., . . . Duijm, L. E. M. (2015). Blinded double reading yields a higher programme sensitivity than non-blinded double reading at digital screening mammography: A prospected population based study in the south of The Netherlands. *European Journal of Cancer (Oxford, England : 1990)*, *51*(3), 391–399. https://doi.org/10.1016/j.ejca.2014.12.008

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, *37*(6), 822–832. https://doi.org/10.1037/0022-3514.37.6.822

Lerner, A. W. (1986). There is more than One Way to be Redundant: A Comparison of Alternatives for the Design and Use of Redundancy in Organizations. *Administration & Society*, *18*(3), 334–359. https://doi.org/10.1177/009539978601800303

Mosier, K. L. [Kathleen L.], Skitka, L. J. [Linda J.], Dunbar, M., & McDonnell, L. (2001). Aircrews and Automation Bias: The Advantages of Teamwork? *The International Journal of Aviation Psychology*, *11*(1), 1–14. https://doi.org/10.1207/S15327108IJAP1101_1

Petty, R. E., Harkins, S. G., Williams, K. D., & Latane, B. (1977). The Effects of Group Size on Cognitive Effort and Evaluation. *Personality and Social Psychology Bulletin*, *3*(4), 579–582. https://doi.org/10.1177/014616727700300406

Poposki, E. M., & Oswald, F. L. (2010). The Multitasking Preference Inventory: Toward an Improved Measure of Individual Differences in Polychronicity. *Human Performance*, *23*(3), 247–264. https://doi.org/10.1080/08959285.2010.487843

Sagan, S. D. (2004). The problem of redundancy problem: Why more nuclear security forces may produce less nuclear security. *Risk Analysis: An Official Publication of the Society for Risk Analysis*, *24*(4), 935–946. https://doi.org/10.1111/j.0272-4332.2004.00495.x

Shepperd, J. A. (1993). Productivity Loss in Performance Groups: A Motivation Analysis. *Psychological Bulletin*, *113*(1), 67–81. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.461.1665&rep=rep1&type=pdf

Skitka, L. J. [L. J.], Mosier, K. L. [K. L.], Burdick, M., & Rosenblatt, B. (2000). Automation bias and errors: Are crews better than individuals? *The International Journal of Aviation Psychology*, *10*(1), 85–97. https://doi.org/10.1207/S15327108IJAP1001_5

Szymanski, K., & Harkins, S. G. (1987). Social loafing and self-evaluation with a social standard. *Journal of Personality and Social Psychology*, *53*(5), 891–897. https://doi.org/10.1037//0022-3514.53.5.891

Taplin, S. H., Rutter, C. M., Elmore, J. G., Seger, D., White, D., Brenner, R. J. [R. James], & Brenner, R. J. [R. J.] (2000). Accuracy of Screening Mammography Using Single Versus Independent Double Interpretation // Accuracy of screening mammography using single versus independent double interpretation. *AJR. American Journal of Roentgenology*, *174*(5), 1257–1262. https://doi.org/10.2214/ajr.174.5.1741257

Thurfjell, E. (1994). Mammography Screening // Screening: One versus Two Views and Independent Double Reading. *Acta Radiologica*, *35*(4), 345–350. https://doi.org/10.1177/028418519403500407

Thurfjell, E., Taube, A., & Tabár, L. (1994). One- versus Two-View Mammography Screening. *Acta Radiologica*, *35*(4), 340–344. https://doi.org/10.1080/02841859409173301

Torka, A.-K., Mazei, J., & Hüffmeier, J. (2021). Together, everyone achieves more-or, less? An interdisciplinary meta-analysis on effort gains and losses in teams. *Psychological Bulletin*, *147*(5), 504–534. https://doi.org/10.1037/bul0000251

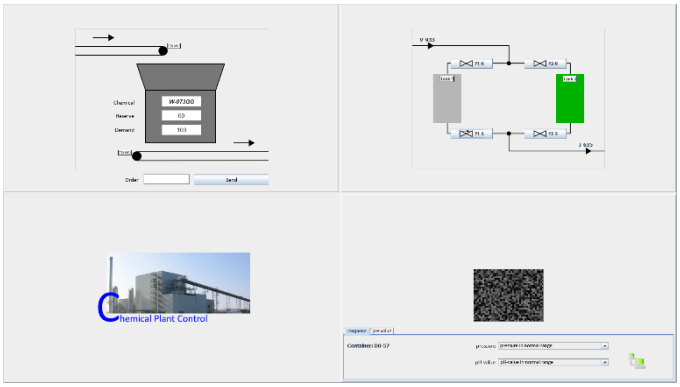

Figure 1. User interface of MTOPS-R used in Experiment 1 (blinded). Upper left: the resource-ordering task (ROT); upper right: the coolant-exchange task (CET); lower right: the quality-control task (QCT) with the network symbol at the bottom right (only present in redundant conditions); lower left: a static image of a chemical plant.

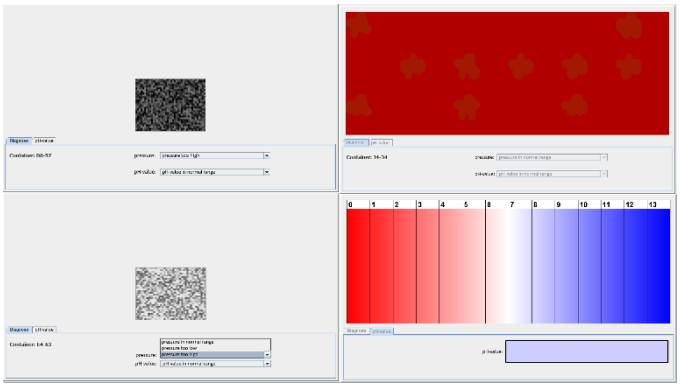

Figure 2. Illustrated interaction stages of the quality-control task (QCT) in the NonR condition. Upper left: clicking on the dark gray pixelated field opens the parameter window for pressure, the pH value window can be accessed by clicking on the corresponding tab labeled “pH value”; upper right: the pressure window with ten dark-red spots (pressure in normal range) is shown; lower right: at the bottom right a color-coded pH value in the normal range is shown, above, the reference scale is depicted (“normal” range: six to ten); lower left: the faded pixelated field is shown as soon as a parameter check has been performed to avoid unintended repeated checks, while below, the selection menu is visible – here with the three options of “normal,” overly high,” and “overly low” pressure.

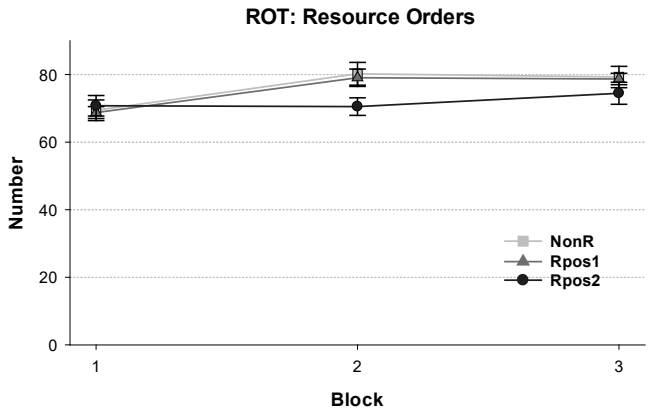

Figure 3. Mean numbers and standard errors of orders conducted per block in the resource-ordering task (ROT) in Experiment 1 (blinded).

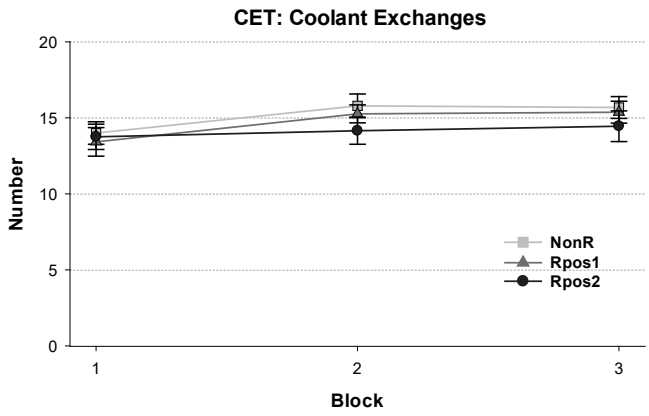

Figure 4. Mean numbers and standard errors of coolant exchanges performed per block in the coolant-exchange tasks (CET) in Experiment 1 (blinded).

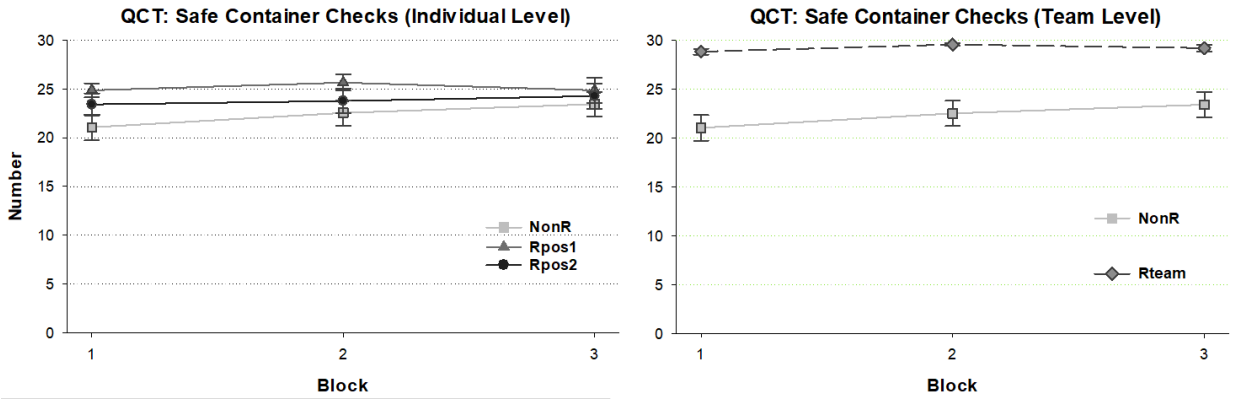

Figure 5. Left: Mean numbers and standard errors of safe container checks in the quality-control task performed per block on the individual level; Right: Mean numbers and standard errors of safe container checks performed per block on the team level in Experiment 1 (blinded).

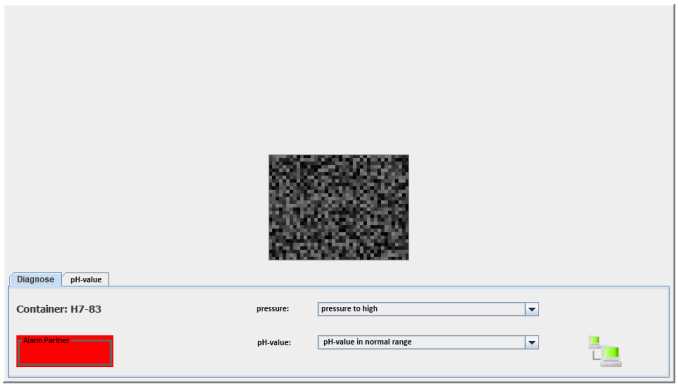

Figure 6. The quality-control task (QCT) in the Rpos2 condition in Experiment 2 (non-blinded). At the bottom left, the red-colored display (partner alarm) indicates that the team partner has found a parameter deviation (green would indicate that no deviation was found by the team partner). A look at the dropdown menu setting indicates that the team partner thinks that the pressure is too high. At the bottom right, a network sign is depicted for all redundant conditions that flashes green when data is transmitted.

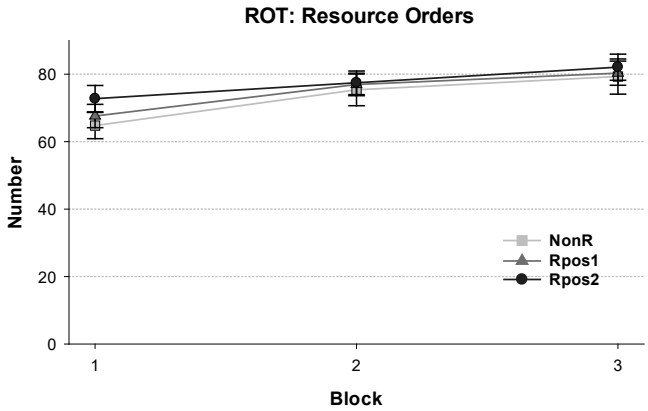

Figure 7. Mean numbers and standard errors of orders conducted per block in the resource-ordering task (ROT) in Experiment 2 (non-blinded).

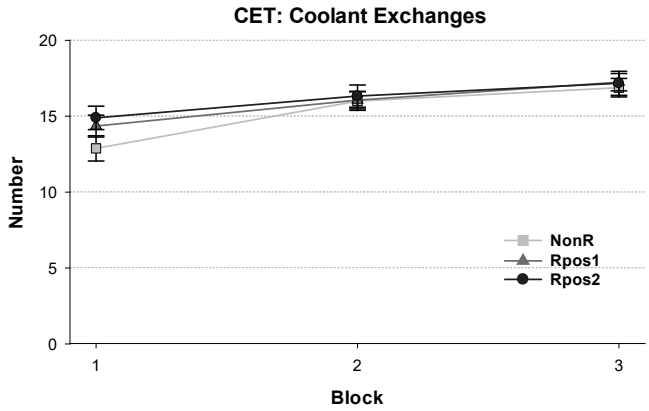

Figure 8. Mean numbers and standard errors of coolant exchanges performed per block in the coolant-exchange tasks (CET) in Experiment 2 (non-blinded).

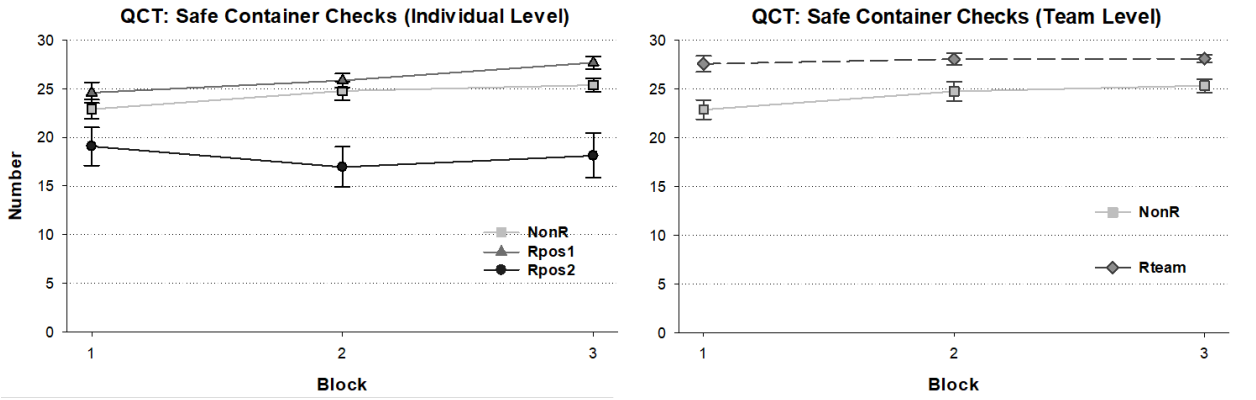

Figure 9. Left: Mean numbers and standard errors of safe container checks in the quality-control task performed per block on the individual level; Right: Mean numbers and standard errors of safe container checks performed per block on the team level in Experiment 2 (non-blinded).

Table 1. Occurrence of unacceptable containers seen in all groups in Experiment 1 (blinded).

Block	1	2	3	(X)
Container	1–30	31–60	61–90	(91–100)
NonR, Rpos1, Rpos2	2, 7, 10, 23, 28, 30	31, 34, 40, 51, 53, 59	61, 65, 69, 81, 82, 90	(94, 96)

Table 2. Means (and standard errors) of subjective effort ratings on a seven-point Likert scale (1 = fully disagree; 7 = fully agree) in Experiment 1 (blinded).

	NonR	Rpos1	Rpos2
#SE1: I made an effort in the QCT.	6.68 (0.12)	6.26 (0.30)	6.25 (0.20)
#SE2: I made as much effort in the QCT as I would have if I had been solely responsible for the task.	-	5.95 (0.35)	5.35 (0.44)
#SE3: Sometimes I relied on the performance of my team partner to have a rest.	-	2.16 (0.39)	2.55 (0.45)

Table 3. Occurrence of unacceptable containers seen in each group with misses (marked with "!") and correct detections (**bold**) of the simulated first checker in Experiment 2 (non-blinded).

Block	1	2	3	(X)
Container	1–30	31–60	61–90	(91–100)
NonR & Rpos1	2, 7, 10, 23, 28, 30	31, 34, 40, 51, 53, 59	61, 65, 69, 81, 82, 90	(/)
Rpos2	**2, 7, 10, 23, 28, 30,**	**31, 34, 40, 51, 53, 59**	**61**, 65!, 69!, **81,** 82!, 90!	**(94, 96)**

Table 4. Means (and standard errors) of subjective effort ratings on a seven-point Likert scale (1 = fully disagree; 7 = fully agree) in Experiment 2 (non-blinded).

	NonR	Rpos1	Rpos2
#SE1: I made an effort in the QCT.	6.41 (0.30)	6.00 (0.35)	5.56 (0.35)
#SE2: I made as much effort in the QCT as I would have if I had been solely responsible for the task. (*)	-	5.71 (0.45)	4.33 (0.51)
#SE3: Sometimes I relied on the performance of my team partner to have a rest. *	-	2.12 (0.35)	3.78 (0.58)