scientific reports

Check for updates

OPEN

The tiny effects of respiratory masks on physiological, subjective, and behavioral measures under mental load in a randomized controlled trial

Robert P. Spang✉ & Kerstin Pieper

Since the outbreak of the coronavirus disease (COVID-19), face coverings are recommended to diminish person-to-person transmission of the SARS-CoV-2 virus. Some public debates concern claims regarding risks caused by wearing face masks, like, e.g., decreased blood oxygen levels and impaired cognitive capabilities. The present, pre-registered study aims to contribute clarity by delivering a direct comparison of wearing an N95 respirator and wearing no face covering. We focused on a demanding situation to show that cognitive efficacy and individual states are equivalent in both conditions. We conducted a randomized-controlled crossover trial with 44 participants. Participants performed the task while wearing an N95 FFR versus wearing none. We measured physiological (blood oxygen saturation and heart rate variability), behavioral (parameters of performance in the task), and subjective (perceived mental load) data to substantiate our assumption as broadly as possible. We analyzed data regarding both statistical equivalence and differences. All of the investigated dimensions showed statistical equivalence given our pre-registered equivalence boundaries. None of the dimensions showed a significant difference between wearing an FFR and not wearing an FFR.

Trial Registration: Preregistered with the Open Science Framework: https://osf.io/c2xp5 (15/11/2020). Retrospectively registered with German Clinical Trials Register: DRKS00024806 (18/03/2021).

Throughout the COVID-19 pandemic, most countries quickly adopted – amongst others – face coverings as a measure to protect the general public. Face coverings can be roughly categorized into face masks (including cloth face coverings), surgical masks, and respirators. According to the FDA, face masks are coverings for the nose and mouth and do not meet filtration efficiency levels (not intended for medical purposes). In contrast, surgical masks meet several protection standards and are considered a medical device. However, their loose fit does not provide complete protection from contaminants[1]. The tight-fitting filtering facepiece respirators (FFRs) such as N95 (US) provide specific filtration efficiencies (at least 95% of small (0.3-micron) particles) and thereby higher virus protection[1,2]. Additionally, surgical masks and FFRs are disposable and should therefore be replaced regularly[1].

Surgical masks and FFRs diminish person-to-person transmission of the SARS-CoV-2 virus[3]. Aerosols better diffuse around one's head by redirecting the exhaled emissions[4]. This process reduces exposures (if other measures such as a sufficient distance are adopted as well)[5,6]. The scientific background at present shows that N95 FFRs without a valve also filter particles, droplets, and aerosols in the in- and exhaled air, which reduces the risk of infection for the person wearing such an FFR, but also, for the people next to them[7] (protection factors of several respirators can be found in[8] and information about filter efficiency in[9]). Modeling the potential for wearing face masks (including homemade cloth masks, surgical masks, and FFRs) demonstrated a drastic decrease in peak hospitalizations and deaths, decreasing the SARS-CoV-2 virus's effective transmission rate[10].

An alarming number of people worldwide question scientific findings and countermeasures against the SARS-CoV-2 virus transmission[11–14]. An early Twitter analysis estimated that around 25% of all tweets regarding the COVID-19 disease contain misinformation[15]. While susceptibility to misinformation seems elevated through social media[16], COVID-19 related misinformation is shared frequently due to failing to question the

Quality and Usability Lab, Institute of Software Engineering and Theoretical Computer Science, Electrical Engineering and Computer Science, Technical University of Berlin, Berlin, Germany. ✉email: spang@tu-berlin.de

content's truthfulness[17]. As such, the potential decline of cognitive performance is discussed. For example, one article concludes that wearing facemasks has physiological and psychological consequences such as—among others—decline in cognitive performance[18]. This is based on a not generalizable finding of declined arterial partial oxygen pressure but unrelated to cognitive performance[19]. However, the manuscript showed several limitations and was, therefore, retracted[20].

Our study aims to provide clarity and evidence against known myths. We investigated multiple dimensions relevant to cognitive performance. We employ a widely acknowledged questionnaire for mental workload (NASA-TLX[21]) as a subjective assessment. The objective measures are physiological values indicating blood oxygen saturation (SpO2) and heart rate variability (HRV). Regarding the behavioral dimensions, we focus on the number of correctly solved problems within the same time interval, the correctness and response times per trial.

Related work. Several studies investigated potential physical consequences or health risks caused by face coverings. Several studies showed that wearing a nonmedical face mask does not lead to a decline in oxygen saturation: in older participants during minimal physical activity[22], no effect on blood and muscle oxygenation in healthy participants[23], not affecting gas exchange during physical activity for neither healthy nor patients with lung function impairment[24], and no change in blood oxygen or the heart rate during rest and a flight simulation of healthy pilots wearing N95 FFRs[25]. There were also no differences in heart rate and blood oxygen parameters in health care workers while a one-hour walk wearing N95 masks[26] and FFR with low filter resistance[27]. However[28], provides evidence for slightly decreased blood oxygen saturation while wearing N95 respirators for very severe COPD patients. Contrarily, only slight differences in heart rate and pulmonary responses were found in[29]. Perceptions of increased body heat most likely originate from warming of the inhaled air, and the facial skin, skin, and core temperature were not affected by wearing an N95 FFR for more than an hour during physical exercise[30].

A subjective evaluation of surgeons reported a hampered performance and increased surgical fatigue while wearing FFP2 masks[31]. Also, a decrease in the blood oxygen saturation and an increase in pulse rates before and after wearing masks[32]. Another study compared wearing an FFR(N95) to exercising without one, which did not show significant differences regarding heart rate, respiratory rate, blood pressure, oxygen saturation, or time to exhaustion in a study by Epstein et al., 2020[33]. Solely end-tidal carbon dioxide (EtCO2) levels were increased while wearing an FFR. Other groups compared the physiological effects of exercising with N95 respirators during pregnancy. Both did not find changed heart rate or blood oxygen levels (although diastolic pressure, mean arterial pressure, and subjective exertion)[34,35].

In an extensive review, several studies investigated the influence of face masks (medical FFR and non-medical face masks) on physiological parameters. They concluded that the effects are negligible and would potentially not impact healthy people even while exercising. However, persons with cardiopulmonary diseases might do experience an effect anyhow[36].

Deliberate misinformation often uses common knowledge to tell an allegedly fact-based story. Some social media accounts connected heavier breathing while wearing FFR with the false claim to reduce blood oxygen saturation. Indeed, respiration behavior (amongst others, frequency and intensity, see[37] for a review) changes while wearing an FFR (especially during exercise), and the physical dead volume of the respiratory system causes breathing to be more strenuous[38]. However, there is no evidence that wearing face masks (cloth/surgical masks or FFR) causes the blood oxygen levels to diminish[22–24,26,27,29]. Nevertheless, the literature lacks investigations tailored to quantify the impact of face masks, especially high filtering N95 FFR, on cognitive performance. We contribute to this research to refute misinformation and face worries regarding a connection between cognitive functioning and wearing N95 FFR.

Regarding our variables of interest, findings from Scholey et al., 1999[39] suggest that in the state of high cognitive demand, the heart rate helps regulate the metabolism, increasing blood oxygen circulation and improving cognitive performance. They showed that oxygen saturation and cognitive performance correlate with each other. Chung et al., 2006[40] presented similar findings where hyperoxic air administration led to increased blood oxygen saturation and improved accuracy in a verbal cognition task compared to regular air administration. In a different study, the HRV was shown to be sensitive for varying levels of cognitive performance. A higher HRV amplitude is suggested to contribute to a decrease in cognitive performance[41]. Mental stress (e.g., induced by mental arithmetic) decreases the HRV, which is suggested to be a regulation process of the autonomic nervous system[42].

Additionally, the HRV seems to be a sensitive indicator to discriminate between rest, physical- and mental load. In a study by Tealman et al., 2011 the combination of a physical task (computer mouse work) and a cognitive task (complex arithmetic) showed a significant decrease in HRV features compared to the physical task alone[43].

The Task Load Index was created to measure demand and the interaction of a subject performing a task[21,44]. It has been frequently used in various like human factors and provides a solid basis for the perceived load.

Behavioral variables are commonly used to measure task difficulty and, thereby, workload. The performance (e.g., measured as a number of solved/correct trials) is expected to decrease when workload reaches a certain threshold[45]. The variation of the difficulty of a task can be indexed in a decrease of correct answers or even in no responses, meaning it was too difficult to solve. At the same time, the duration for producing a response increases if the task is more complex and thereby more mentally demanding than the one before[46].

Our contribution. Given the body of evidence, we hypothesize equivalence of blood oxygen saturation while wearing an N95 FFR compared to not wearing one. Further, we hypothesize equivalence of the cognitive demand of the FFR and the no-FFR condition. We expect that the participants perform equally well in both test conditions. In terms of behavioral data, we hypothesize equivalence between the conditions regarding the num-

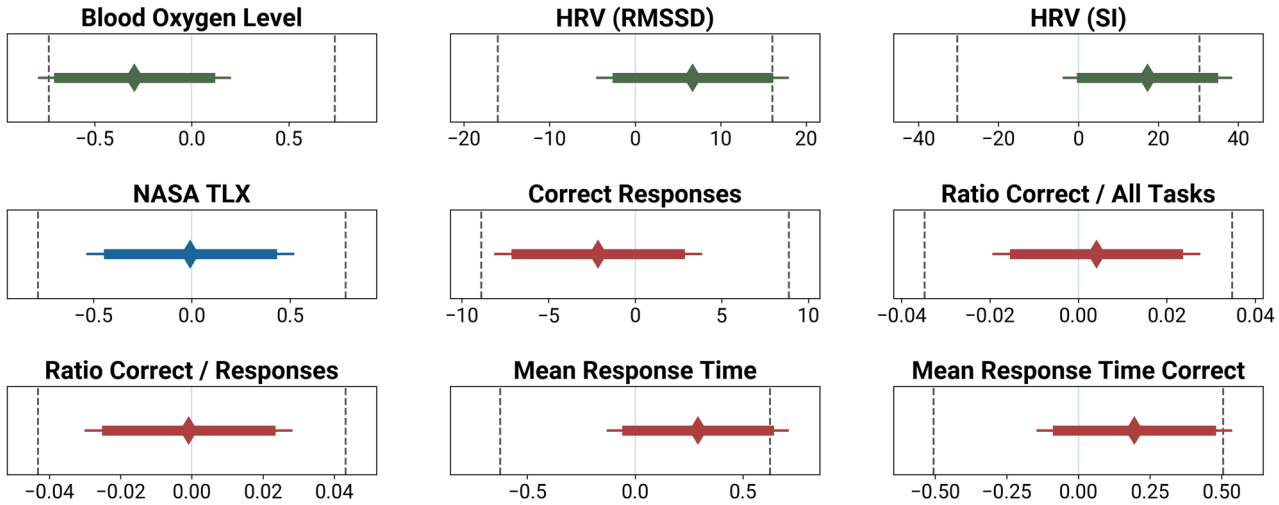

Figure 1. Equivalence boundaries (dotted lines left and right), mean of the mask / no-mask difference (diamond) and the 95% confidence interval (thin line; for the null hypothesis significance test), as well as the 90% confidence interval for the TOST (thick line). The x-axis shows the mean difference in the unit of the metric.

ber of correctly solved tasks, the ratio of correct responses to all tasks presented, the ratio of correct responses to all responses given, the average response time, and the average response time of correct responses.

In terms of physiological data, we assume similarbehavior in both test conditions. The task to be performed has a cognitive focus and is carried out under time pressure. We expect no physical exertion in the relaxed sitting position. Thus, only cognitive demand could influence the physiological parameters as described in the mentioned literature. Providing that cognitive demand is equal in both conditions (with and without an FFR), we hypothesize equivalent results regarding participants' HRV, SpO2, TLX scores, and task performance. This study adheres to CONSORT guidelines.

Results

For all following Two One-Sided Test of Equivalence (TOST) procedures, we employed equivalence boundaries of $d_z = \pm 0.45$. This smallest effect size of interest (SESOI) translates to the absolute values of the equivalence boundaries reported in the following paragraphs. Figure 1 provides an overview of the TOST confidence intervals and the null hypotheses significance tests, together with the equivalence boundaries.

For both HRV analyses, we had to exclude three datasets due to incomplete recordings. Hence, both are based on data from 41 participants. Given the chosen alpha level of $\alpha = 0.05$ and the pre-defined equivalence bounds of $d_z = \pm 0.45$, both HRV TOST results have a statistical power of $1-\beta = 0.78$. All other tests are based on all 44 participants, resulting in statistical power of the TOSTs of $1-\beta = 0.82$.

Physiological data. See Fig. 2 for a visualization of the blood oxygen saturation and the HRV measurement (RMSSD) per condition. All result graphs share the same format and visualize different aspects of the group comparison. First, we contrast the distribution of the two groups. For a precise understanding about outliers, centers and spread of the inner 50%, we then align box-plots. In addition to that, we underline the equality of the group means by adding simple bar-plots with 95%-range whiskers.

Physiological: blood oxygen levels. The mean difference of blood oxygen level between wearing an FFR (95% CI: 96.04–97.64%) and not doing so (95% CI: 96.48–97.79%) immediately after performing the 15 min of mental calculation is − 0.3% (difference Median: 0%, IQR: 2%). The increase of blood oxygen level without a mask has a negligible effect size of $d_z = -0.12$. A Shapiro–Wilk test indicated a violation of the assumption of normality ($W = 0.92$, $p = 0.004$). Hence, we employed a robust TOST procedure using Wilcoxon signed-rank test. To compare the measurements of two conditions, we define an equivalence interval. It is derived from our pre-defined effect size of $d_z = \pm 0.45$, which translates to ±0.736 in the units of the metric at hand (percent in this case). Hence, the lower equivalence boundary $\Delta_L = -0.74\%$ and the upper equivalence boundary $\Delta_U = 0.74\%$. The TOST procedure reveals that the effect observed is statistically equivalent; the larger of the two p values is less than $\alpha = 0.05$ ($V = 682$, $p = 0.014$). According to the Neyman-Pearson approach, this means that one can reject the hypothesis that the true effect is greater than $d_z = \pm 0.45$ and act as if the effect size falls within these equivalence bounds[47]. According to our pre-registration, we additionally run an exploratory null hypothesis significance test. A pairwise Wilcoxon signed-rank test returned nonsignificant ($V = 154$, $p = 0.259$). Hence the H0 of no difference between groups is not rejected.

Physiological: heart rate variability (RMSSD). The mean difference of RMSSD between wearing an FFR (95% CI: 28.81–58.6 ms) and not wearing one (95% CI: 29.27–44.69 ms) in the last five minutes of each condition is 6.73 ms (difference Median: 1.63 ms, IQR: 11.92 ms). The decrease of the RMSSD without a mask has a negli-

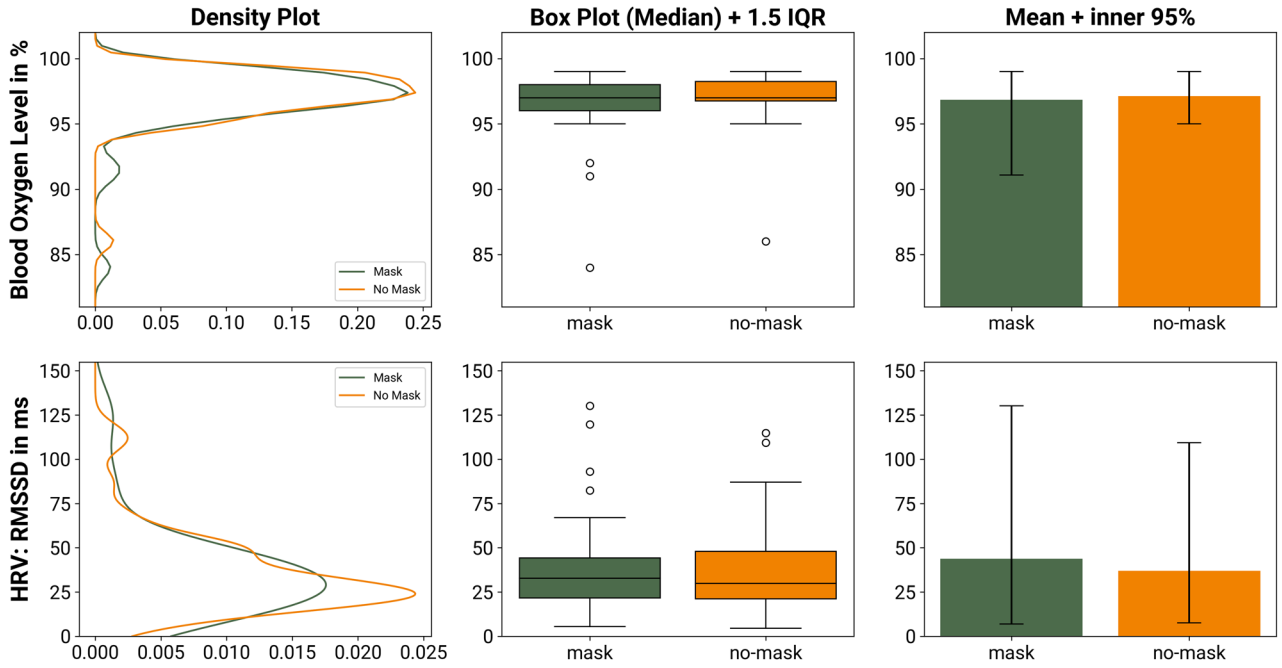

Figure 2. Comparison of the physiological metrics (blood oxygen level and HRV) while wearing an FFR and not wearing an FFR. The density plots to the left describe the similarity of the distributions of the two groups. The box-plots in the center column compare the median and the interquartile range (IQR) and provide an assessment of potential outliers. The bar charts to the right compare the plain mean of the two group; the whiskers depict the inner 95% of the recorded data.

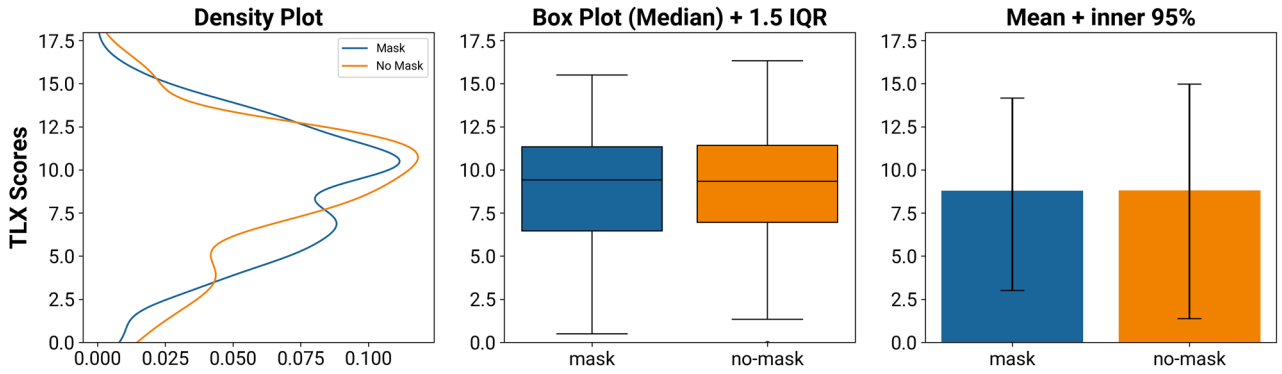

Figure 3. Comparison of the subjective load ratings (NASA TLX) while wearing an FFR and not wearing an FFR. While the distribution reveals minor differences between the groups, these are averaged out when comparing mean and median values.

gible effect size of $d_z = 0.15$. A Shapiro–Wilk test indicated a violation of the assumption of normality ($W = 0.38$, $p < 0.001$), hence we employed a robust TOST procedure using Wilcoxon signed-rank test, with equivalence bounds of $\Delta_L = -16.06$ ms and $\Delta_U = 16.06$ ms. It reveals that the effect observed is statistically equivalent, the larger of the two p values is less than $\alpha = 0.05$ ($V = 56$, $p < 0.001$). We additionally ran an exploratory null hypothesis significance test. A pairwise Wilcoxon signed-rank test returned nonsignificant ($V = 519$, $p = 0.257$).

Subjective data. To investigate the NASA-TLX scores, we first computed the difference between post-task and baseline ratings. The mean difference of these scores is −0.01 (difference Median: 0.1, IQR: 2.1, see Fig. 3). The decrease of the TLX score without an FFR (95% CI: 7.66–9.95) has a negligible effect size of $d_z = -0.002$ (95% CI of the mask condition: 7.74–9.85). The assumption of a normal distribution was not rejected ($W = 0.99$, $p = 0.919$), so we used a TOST procedure based on Welch's paired t-test with equivalence bounds $\Delta_L = -0.78$ and $\Delta_U = 0.78$. It reveals that the effect observed is statistically equivalent, the larger of the two p values is less than $\alpha = 0.05$ ($t(43) = 2.96$, $p = 0.003$). An exploratory null hypothesis significance test (pairwise Welch's t-test) returned nonsignificant ($t(43) = -0.03$, $p = 0.977$).

Behavioral data. See Fig. 4 for a visualization of the following five behavioral performance data per condition.

Behavioral: correct responses. The mean difference between the number of correct responses while wearing an FFR (95% CI: 79.13–97.37) against while not wearing one (95% CI: 82.38–98.39) is −2.14 (difference Median: 3.5, IQR: 24.5). The increase of correct responses in conditions without an FFR has a negligible effect size of $d_z = -0.08$. The assumption of a normal distribution was rejected (Shapiro–Wilk test, $W = 0.94$, $p = 0.015$), so we used a robust TOST procedure based around the Wilcoxon signed-rank test with equivalence bounds of $\Delta_L = -8.89$ and $\Delta_U = 8.89$. It reveals that the effect observed is statistically equivalent ($V = 680$, $p = 0.016$). An exploratory null hypothesis significance test (pairwise Wilcoxon signed-rank test) returned nonsignificant ($V = 496$, $p = 0.995$).

Behavioral: ratio correct responses/all tasks. We investigate the ratio of correct responses against the number of all responses given (correct and incorrect). The mean difference between an FFR and no FFR is nearly zero (difference Median: − 0.01, IQR: 0.09). The effect induced by the FFR (95% CI: 0.55–0.64) is negligible ($d_z = 0.03$, 95% CI of the no-FFR condition: 0.55–0.63. The assumption of a normal distribution was not rejected ($W = 0.96$, $p = 0.133$), so we used a TOST procedure based on Welch's paired t-test with equivalence bounds $\Delta_L = -0.04$ and $\Delta_U = 0.04$. It reveals that the effect observed is statistically equivalent, the larger of the two p values is less than $\alpha = 0.05$ ($t(43) = -2.64$, $p = 0.005$). An exploratory null hypothesis significance test (pairwise Welch's t-test) returned nonsignificant ($t(43) = 0.35$, $p = 0.728$).

Behavioral: ratio correct responses/responses given. Next, we investigate the ratio of correct responses against the number of all tasks presented.

The mean difference between FFR and no FFR is nearly zero (difference Median: − 0.01, IQR: 0,1). The effect induced by the FFR (95% CI: 0.67–0.78) is negligible ($d_z = -0.01$, 95% CI of the no-FFR condition: 0.68–0.78. The assumption of a normal distribution was not rejected ($W = 0.98$, $p = 0.524$), so we used a TOST procedure based around Welch's paired t-test with equivalence bounds of $\Delta_L = -0.04$ and $\Delta_U = 0.04$. It reveals that the effect observed is statistically equivalent, the larger of the two p values is less than $\alpha = 0.05$ ($t(43) = 2.92$, $p = 0.003$). An exploratory null hypothesis significance test (pairwise Welch's t-test) returned nonsignificant ($t(43) = -0.06$, $p = 0.950$).

Behavioral: mean response time. The mean difference between a mask and no mask of the average response time is 0.29 s (difference Median: − 0.05 s, IQR: 1.46 s). The decrease of the response time in conditions without an FFR (95% CI: 5.06–5.82 s) has a small effect size of $d_z = 0.21$ (95% CI of the FFR condition: 5.29–6.17 s). The assumption of a normal distribution was rejected (Shapiro–Wilk test, $W = 0.91$, $p = 0.002$), so we used a robust TOST procedure based around the Wilcoxon signed-rank test with equivalence bounds of $\Delta_L = -0.63$ s and $\Delta_U = 0.63$ s. It reveals that the effect observed is statistically equivalent ($V = 329$, $p = 0.026$). An exploratory null hypothesis significance test (pairwise Wilcoxon signed-rank test) returned nonsignificant ($V = 529$, $p = 0.699$).

Behavioral: mean response time of correct responses. Lastly, we investigate the average response time of only correct responses. The mean difference between an FFR (95% CI: 4.9–5.6 s) and no FFR (95% CI: 4.76–5.35 s) is 0.2 s (difference Median: − 0.03 s, IQR: 0,93 s). The decrease of the response time in conditions without an FFR has a negligible effect size of $d_z = 0.18$. The assumption of a normal distribution was rejected (Shapiro–Wilk test, $W = 0.93$, $p = 0.009$), so we used a robust TOST procedure based around Welch's paired t-test with equivalence bounds of $\Delta_L = -0.51$ s and $\Delta_U = 0.51$ s. It reveals that the effect observed is statistically equivalent ($t(43) = -1.83$, $p = 0.037$). An exploratory null hypothesis significance test (pairwise Welch's t-test) returned nonsignificant ($t(43) = 1.15$, $p = 0.255$).

Discussion

The blood oxygen saturation shows a slight decrease of 0.3% after wearing an FFR. This effect is statistically insignificant. Although some discussions against the use of facial masks argue that FFR would impair the body's oxygen supply, this is unstrained by our findings. Instead, we found statistical equivalence and no difference between the test conditions. The HRV metric (RMSSD) showed statistical equivalence when comparing the FFR against the no-mask condition and no significant difference from each other. The HRV seems to decrease slightly (statistically insignificant) in the no-FFR condition on a descriptive level.

When interpreting the HRV metrics as mental load indicators, the RMSSD typically drops if the participant is more strained[48]. On a descriptive level, we find opposing results: the RMSSD indicates slightly more strain, higher intensity load, and focus in the no-FFR condition. This underlines that the changes induced by the FFR cause less variability than the HRV can interpret reasonably.

The subjective NASA-TLX ratings show that the participants perceived a statistically equivalent workload between wearing an FFR and not wearing one. This result may come as a surprise: Because we did not include a blinding protocol, participants were always fully aware of wearing an FFR and not. We did not explicitly tell them about our research question before the experiment was over. However, some participants might have figured out why to wear an FFR sometimes and why not (none of the participants implied so). Nevertheless, because we cannot rule out the possibility of the participants guessing our research question and perhaps even being biased towards governmental pandemic restrictions, it remains possible to have recorded biased results. For this very reason, it seems remarkable that the subjective TLX ratings show no evidence of favoring one of

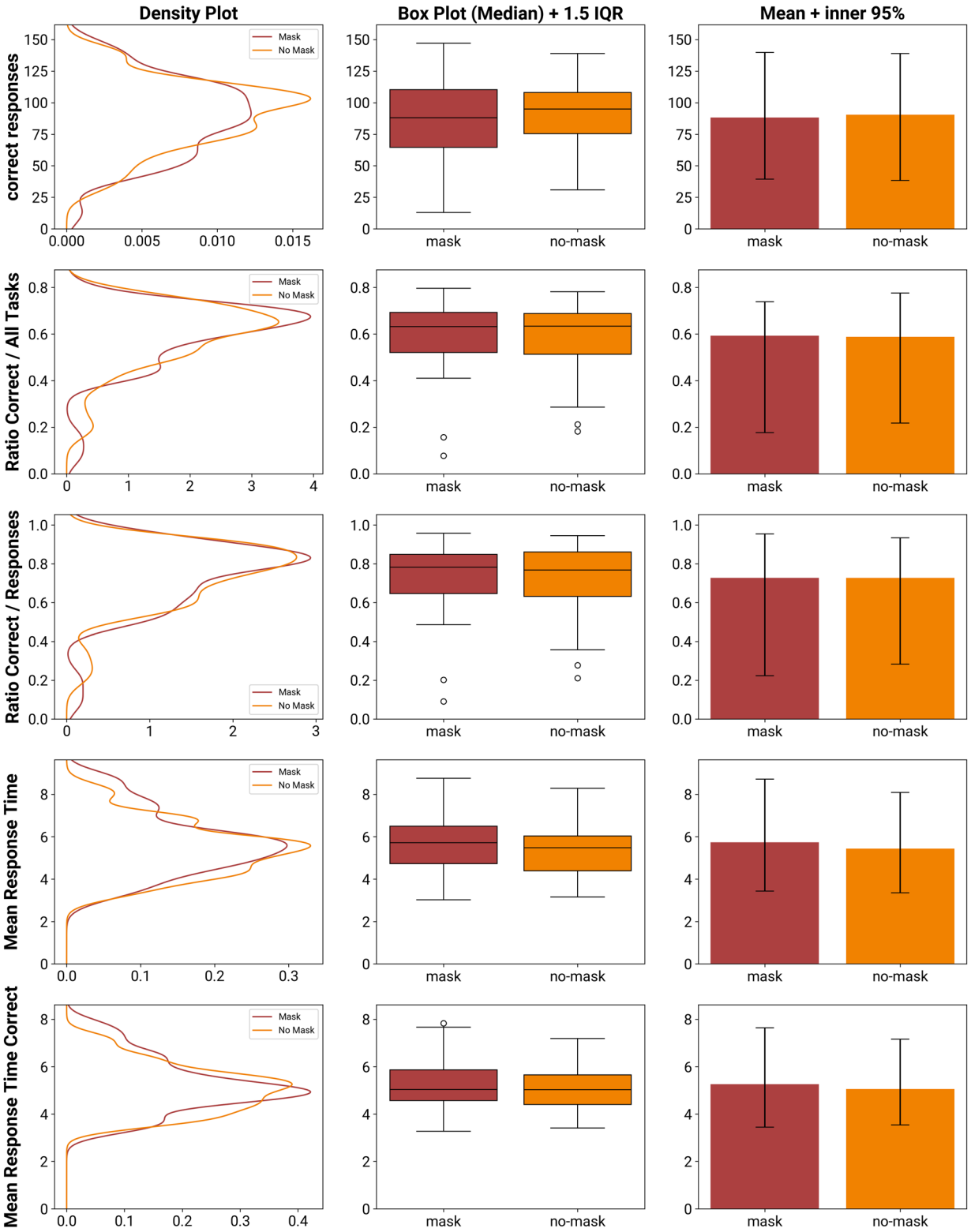

Figure 4. Comparison of the behavioral measures while wearing a mask and not wearing a mask. The dimensions compared are the absolute number of correct responses, the ratios of correct responses against all tasks presented as well as against the number of responses, the mean response time per task, and the mean response time of only the correct responses.

the conditions, not even on a descriptive level. Mainly since the subjective assessment includes an item asking for physical demand that might capture aspects such as wearing comfort (e.g.,[49] reported "marked discomfort" of the participants wearing FFP2 masks, although the study, in general, is heavily debated, see[50,51]. Other subjective reports also mentioned comfort limitations, e.g., discomfort if the wearer has facial hair[52] or the problem of subjective difficulty to breathe[53]). Therefore, it seems to be an even greater confirmation that wearing an FFR does not limit the wearer's performance. We deem it unlikely to confound all our different metrics regarding the possible condition awareness, primarily since we investigated a broad spectrum of varying measurement dimensions.

Regarding the behavioral data, the FFR's influence reached a small effect ($d_z = 0.214$) for the mean response time; for all other parameters, the effect was negligible. However, this small effect is a statistical artifact that could not be shown to cause a statistical difference. Moreover, the variability induced by the FFR is equivalent to the variability of not wearing one (given our pre-defined equivalence bounds). This means that neither did the participants solve more tasks in 15 min when not wearing an FFR, nor was their correct response ratio any better. Even the average response time was statistically equivalent to the FFR condition. Hence, we deem these findings to refute the claim that facial masks potentially reduce cognitive performance in a meaningful magnitude.

While the participants sat alone in the lab room, not wearing an FFR in one condition, we decided to provide them with N95 FFR (CE-certified FFP2 in Europe / KN95 in China) for the FFR condition. FFRs generally sit tighter on the face, suffer from less face seal leakage, and its filter medium offers more substantial filter characteristics than surgical masks[54]. Since most homemade masks have even less powerful filtering properties than surgical masks[55], two interpretations can be drawn for wearers of these more superficial masks: Either the FFR itself primarily attributes the effects found. Then one could suggest that the impact would be diminished even further when wearing surgical or homemade masks. Alternatively, the observed effects are simple non-systematic measurement artifacts. In this case, one would observe effects in the same order of magnitude and similar variations when replicating our work with surgical and homemade masks. In either case, degradation of cognitive performance is not to be expected from wearing FFRs.

The equivalence boundaries we chose are smaller than the effects reported so far. However, this assessment is somewhat rough since no previous work that we are aware of investigated similar relationships and the reported effect sizes of[33,56] hat to be converted to standardized Cohen's d. To account for conversion errors, we defined our threshold slightly below the definition of a large effect size (which would be $d_z = 0.5$). We decided to do this because it compromises meaningfulness and a realistic number of test participants. Nevertheless, this definition is potentially our Achilles' heel: our statistical tests' significance relies heavily on the equivalence boundaries. One could argue that these are just wide enough for all our equivalence tests to turn out significantly. While this is de facto the case, it is essential to point out that we pre-defined our equivalence boundaries in the preregistration before assessing the recorded data and before most of the data has been sampled (as recommended by[47]). However, a replication with smaller equivalence boundaries and a larger sample size would further substantiate our findings.

Other than that, it is worth pointing out that our mental load condition lasted only 15 min. In discussions with mask-skeptic people, we heard the argument that wearing masks for a whole day would impact cognitive functioning. Our comparison based around two 15 min conditions cannot easily be compared to a whole day. However, it is known from the literature that the time in which a change of inhaled air is reflected in blood oxygen readings lies within several seconds up to a minute (e.g.[57]. Hence, if there is no evidence for any impact of the masks after wearing them for 15 min, there is little reason to believe that this drastically changes after several hours.

Methods

Task. The main task consisted of solving basic arithmetic equations (addition, subtraction, multiplication, or division) presented visually. Each equation was composed of two numbers (1 to 3 digits) and one operator. All results were positive integers. We decided to implement mental arithmetic because these tasks are suitable for inducing cognitive processing[58], and the task allows usto vary difficulty levels of the task[59,60]. Additionally, this stimulus is suited to simulate office work[43,61].

The response time for typing in the correct arithmetic solution was limited, depending on the estimated difficulty of the task. This estimation was done using a prediction model based around Thomas' Q-value, 1963[62] to estimate a primary arithmetic task's difficulty. This task design allowed us to induce a constant, high mental load for each condition's duration.

Procedure. We experimented in a small and bright lab room at the Technical University of Berlin during regular office hours. The pandemic situation forced us to limit the time spent with the participants to < 15 min. Hence, the general introduction was done in a separate room, and the time together was spent instructing the conditions. Before and after each experiment, the lab room was heavily ventilated, and all surfaces and devices were disinfected. Due to the strict regulations, the entire floor was hardly occupied, which guaranteed a quiet environment.

The participants were equipped with the chest strap (model: Polar H10; Polar Electro Oy). The ECG data (HR and RR-intervals) was recorded directly via Polar's Bluetooth API to a dedicated smartphone running our recorder app. Additionally, they wore a Comtec Pulse Oximeter, model CMS 50D, which we put on the participant's index finger of the non-dominant hand. The device was active throughout the whole experiment. The FFR we provided was unvented FFP2 NR N95 / KN95 (model number: B13086; Samding Craftwork Co., LTD, Jinniu Daojiao Dongguan Guangdong, China) with a full CE certification (CE 2163, EN 149: 2001 + A1: 2009). Every provided FFR came in a standard size. Since all subjects were adults, we did not use a custom FFR size. However, we instructed the subjects to put on the mask in a well-fitting manner via the adjustable nose clip.

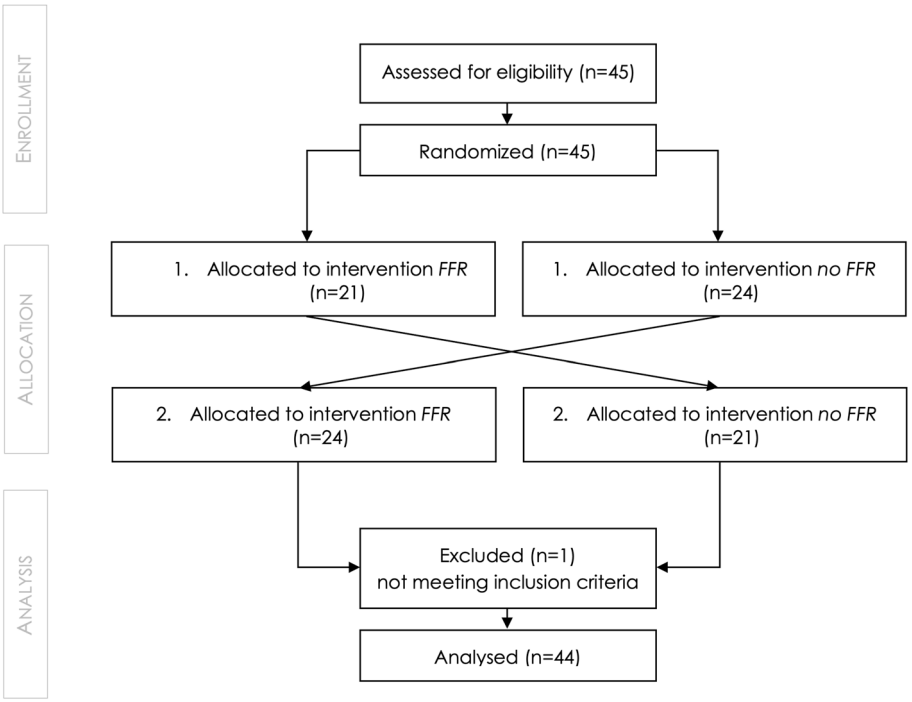

Figure 5. Flow-Chart of the participant flow through the data acquisition.

The stimulus presentation was done via smartphone (iPhone XR; Apple Inc.; 6′1-inch screen).

The stimulus presentation was implemented as an iOS app. From top to bottom the app displayed the task, the time left together with a shrinking progress bar, and a key pad to enter the solution and two buttons to delete and to confirm the response.

The measurement of the pulse oximeter is continuous. Therefore, we instructed the participants to note down the current reading of the oximeter right after each test condition.

At the start of the experiment, we conducted a baseline measurement of HRV and subjective data. Therefore, participants filled in a NASA-TLX rating in which they rated their current situation (e.g., waiting in the foyer). Additionally, the participants performed their first blood oxygen measurement as the experimenter showed them beforehand. The baseline recording was also a practice to do all the measurements correctly and took about 5 min in total.

Each participant was assigned to a random order of the conditions on arrival (see Fig. 5). This randomization was not known to the experimenters before the start of the data collection. The study app announced which condition came next (and logged this to a log file). We used the Swift 5 standard library to generate a random order of the conditions. In one test condition, they performed the arithmetic calculations while wearing a mask, in the other without a mask. Both test conditions had a duration of 15 min. After each condition, the participants had to fill in the NASA-TLX ratings and ran the blood oxygen measurement manually.

After the completion of both conditions, including both post measurements, the experimenter removed the sensors. The participants confirmed the monetary compensation with a receipt.

Participants. The conduction of the study took place between October and November 2020. An a priori power analysis for matched pairs TOST (alpha = 0.05, power = 0.8, equivalence bound dz = ± 0.45, cf.[63]) resulted in a minimum required number of 43 participants. We recruited 45 participants to account for possible exclusions due to a lack of correct responses. Twenty-four of the 45 participants identified themselves as female. The mean age was 30.3 years (Median: 29y, IQR: 8y, ranging from 20 to 64y). Twenty-four of the participants hold at least one academic degree; Twenty-two participants were currently enrolled, students. The majority of participants were recruited via the university participant database. It ensures that the offered studies are only visible to people who match predefined criteria, so (usually) no one has to be excluded later on. The only criteria we employ are being aged between 18 and 65, being fluent in German, and having normal or corrected to normal vision. Participants got a monetary compensation of a fixed amount of 12 Euro plus a performance-dependent addition of up to 6 Euro. Besides, some colleagues declared themselves willing to participate. The study protocol was approved by the ethics committee of Technical University Berlin, Faculty IV Electrical Engineering and Computer Science (ethics ID: FT_2020_11). The conductance of the experiment was according to the declaration of Helsinki. The participants obtained informed consent in written form and declared their agreement with the procedure by signature before the recordings began.

The study was a randomized controlled trial. We employed a crossover study design with 45 participants (see Fig. 5). All participants were unaware of the conditions and the differences that we are investigating. However,

since they are being told to wear or not wear an FFR before a condition started, they potentially could guess the mask itself is a manipulated factor. One participant had to be excluded from our dataset due to our exclusion criterion defined in the pre-registration. The subject did not reach a minimum performance of min 10% correct trials in the task, which was necessary to be considered in the analysis. So, we considered 44 participants in our general analysis.

Statistical analysis. Equivalence tests examine whether the presence of large enough effects to be considered meaningful can be rejected[47]. This TOST procedure compares an observed distribution against the boundaries of a predefined equivalence interval. The statistical procedure is then identical to two one-sides t-tests (or equivalent) for determining if the distribution at hand is significantly below the upper equivalence interval boundary, and if it is as well significantly above the lower equivalence boundary. The procedure is thoroughly described by Lakens and colleagues[47].

In our case, the equivalence test is used to examine whether the difference between wearing an FFR and not wearing one is at least as extreme as a mid-sized effect of $d_z = \pm 0.45$.

We defined the SESOI as $d_z = \pm 0.45$, based on analyzing reported effect sizes of the related literature (especially[33,56]). However, previous studies had a slightly different focus, so we assumed a slightly smaller effect size than what the colleagues reported. Hence, we decided to choose a fixed effect size just below the "large effect" -guideline $d_z = \pm 0.5$. Our definition of the SESOI was part of our pre-registration.

Depending on whether the data is normally distributed, we employ a TOST procedure based on Welch's t-test ("TOSTER" package v0.3.4 for R), or on the Wilcoxon signed-rank test with continuity correction ("stats" package v3.5.1 of R).

Regarding the HRV measures we binned the time span of each condition into five minutes intervals. For the statistical analysis we computed RMSSD and SI measures for the last interval of each condition only. This way, variations in HRV induced by the onset of each condition should be diminished.

Conclusion

We hypothesized that wearing an FFR while performing a demanding, cognitive task for 15 min does not statistically differ from completing the same task without an FFR. To do so, we created a testbed allowing us to measure physiological changes in blood oxygen level and heart rate variability, subjective assessment of the mental load, and behavioral performance data. All our findings support all our hypotheses. All metrics recorded with an FFP2 mask are statistically equivalent to not wearing a mask, given our pre-defined equivalence interval of $d_z = \pm 0.45$. We interpreted that we can reject the hypothesis of a large effect induced by an FFR (larger than $d_z = 0.45$). In addition to the statistical equivalence test, we did not find any statistical differences between the two groups. We provided a direct comparison between wearing an FFR and not wearing one. The combination of physiological, subjective, and behavioral data delivers a measurement tool that allows us to detect potential differences objectively and subjectively. Out of that, we are confident that our results support previous research findings and deliver valuable contributions, especially in terms of the current mask debate.

Data availability

The datasets generated during and/or analyzed during the current study, as well as the analysis scripts themselves, are available in the Open Science Framework repository: https://osf.io/c2xp5.

Received: 2 December 2020; Accepted: 20 September 2021
Published online: 01 October 2021

References

1. Health, C. for D. and R. Face Masks, Including Surgical Masks, and Respirators for COVID-19. *FDA* (2021).
2. Respirator FAQs | NPPTL | NIOSH | CDC. https://www.cdc.gov/niosh/npptl/topics/respirators/disp_part/respsource3basic.html (2021).
3. Leung, N. H. L. *et al.* Respiratory virus shedding in exhaled breath and efficacy of face masks. *Nat. Med.* **26**, 676–680 (2020).
4. Asadi, S. *et al.* Efficacy of masks and face coverings in controlling outward aerosol particle emission from expiratory activities. *Sci. Rep.* **10**, 15665 (2020).
5. Li, Y. *et al.* Role of ventilation in airborne transmission of infectious agents in the built environment-a multidisciplinary systematic review. *Indoor Air* **17**, 2–18 (2007).
6. Morawska, L. & Milton, D. K. It is time to address airborne transmission of COVID-19. *Clin. Infect. Dis.* **6**, ciaa939 (2020).
7. Sommerstein, R. *et al.* Risk of SARS-CoV-2 transmission by aerosols, the rational use of masks, and protection of healthcare workers from COVID-19. *Antimicrob. Resist. Infect. Control* **9**, 100 (2020).
8. National Institute for Occupational Safety and Health (NIOSH). Respirator Selection Logic: NIOSH (2004).
9. Sbihi, H., Nicas, M., Rideout KJBCfDC, Vancouver BC. Evidence Review: Using masks to protect public health during wildfire smoke events (2014).
10. Eikenberry, S. E. *et al.* To mask or not to mask: Modeling the potential for face mask use by the general public to curtail the COVID-19 pandemic. *Infect. Dis. Model.* (2020).
11. Evanega, S., Lynas, M., Adams, J., Smolenyak, K. & Insights, C. G. *Coronavirus misinformation: Quantifying sources and themes in the COVID-19 'infodemic'.* (The Cornell Alliance for Science, 2020).
12. Mian, A. & Khan, S. Coronavirus: The spread of misinformation. *BMC Med.* **18**, 1–2 (2020).
13. Sharma, K. *et al.* Coronavirus on social media: Analyzing misinformation in Twitter conversations. ArXiv Preprint. ArXiv200312309 (2020).
14. Tasnim, S., Hossain, M. M. & Mazumder, H. Impact of rumors or misinformation on coronavirus disease (COVID-19) in social media (2020).
15. Kouzy, R. *et al.* Coronavirus goes viral: Quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus* **12**, 2020 (2020).
16. Roozenbeek, J. *et al.* Susceptibility to misinformation about COVID-19 around the world. *R. Soc. Open Sci.* **7**, 201199 (2020).

17. Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G. & Rand, D. G. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychol. Sci.* **31**, 770–780 (2020).
18. Vainshelboim, B. Facemasks in the COVID-19 era: A health hypothesis. *Med. Hypotheses* **146**, 110411. https://doi.org/10.1016/j.mehy.2020.110411 (2021).
19. Tw, K. *et al.* The physiological impact of wearing an N95 mask during hemodialysis as a precaution against SARS in patients with end-stage renal disease. *J. Formos. Med. Assoc.* **103**(8), 624–628 (2004).
20. Vainshelboim, B. Retraction notice to 'Facemasks in the COVID-19 era: A health hypothesis' [Medical Hypotheses 146 (2021) 5]. *Med. Hypotheses* (2021). https://doi.org/10.1016/j.mehy.2021.110601.
21. Hart, S. G. (2006, October). NASA-task load index (NASA-TLX); 20 years later. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 50, No. 9, pp. 904–908). Sage CA: Los Angeles, CA: Sage publications.
22. Chan, N. C., Li, K. & Hirsh, J. Peripheral oxygen saturation in older persons wearing nonmedical face masks in community settings. *JAMA* **324**, 2323–2324 (2020).
23. Shaw, K., Butcher, S., Ko, J., Zello, G. A. & Chilibeck, P. D. Wearing of cloth or disposable surgical face masks has no effect on vigorous exercise performance in healthy individuals. *Int. J. Environ. Res. Public Health* **17**, 8110 (2020).
24. Samannan, R., Holt, G., Calderon-Candelario, R., Mirsaeidi, M. & Campos, M. Effect of face masks on gas exchange in healthy persons and patients with COPD. *Ann. Am. Thorac. Soc.* https://doi.org/10.1513/AnnalsATS.202007-812RL (2020).
25. Sammito, S., Müller, G. P. J., Erley, O. M. & Werner, A. Impact of in-flight use of FFP2 masks on oxygen saturation: An experimental crossover study. *J Travel Med* https://doi.org/10.1093/jtm/taab018 (2021).
26. Roberge, R. J., Coca, A., Williams, W. J., Powell, J. B. & Palmiero, A. J. Physiological impact of the N95 filtering facepiece respirator on healthcare workers. *Respir Care.* **55**(5), 569–577 (2010).
27. Roberge, R. J. *et al.* Impact of low filter resistances on subjective and physiological responses to filtering facepiece respirators. *PLoS ONE* **8**(12), e84901 (2013).
28. Kyung, S. Y., Kim, Y., Hwang, H., Park, J. W. & Jeong, S. H. Risks of N95 face mask use in subjects with COPD. *Respir. Care* **65**(5), 658–664 (2020).
29. Kim, J. H., Benson, S. M. & Roberge, R. J. Pulmonary and heart rate responses to wearing N95 filtering facepiece respirators. *Am. J. Infect. Control* **41**(1), 24–27 (2013).
30. Roberge, R., Benson, S. & Kim, J. H. Thermal burden of N95 filtering facepiece respirators. *Ann. Occup. Hyg.* **56**(7), 808–814 (2012).
31. Yánez Benítez, C. *et al.* Impact of personal protective equipment on surgical performance during the COVID-19 pandemic. *World J. Surg.* **44**, 2842–2847 (2020).
32. Beder, A., Büyükkoçak, Ü., Sabuncuoğlu, H., Keskil, Z. A. & Keskil, S. Preliminary report on surgical mask induced deoxygenation during major surgery. *Neurocirugía* **19**, 121–126 (2008).
33. Epstein, D. *et al.* Return to training in the COVID-19 era: The physiological effects of face masks during exercise. *Scand. J. Med. Sci. Sports* (2020).
34. Kim, J. H., Roberge, R. J. & Powell, J. B. Effect of external airflow resistive load on postural and exercise-associated cardiovascular and pulmonary responses in pregnancy: A case control study. *BMC Pregnancy Childbirth* **15**(1), 1–8 (2015).
35. Roberge, R. J., Kim, J. H. & Powell, J. B. N95 respirator use during advanced pregnancy. *Am. J. Infect. Control* **42**(10), 1097–1100 (2014).
36. Hopkins, S. R. *et al.* Face masks and the cardiorespiratory response to physical activity in health and disease. *Ann. Am. Thoracic Soc.* **18**(3), 399–407 (2021).
37. Johnson, A. T. Respirator masks protect health but impact performance: A review. *J. Biol. Eng.* **10**(1), 1–12 (2016).
38. Johnson, A. T. *et al.* Effect of external dead volume on performance while wearing a respirator. *Am. Ind. Hygiene Assoc.* **61**(5), 678–684 (2000).
39. Scholey, A. B., Moss, M. C., Neave, N. & Wesnes, K. Cognitive performance, hyperoxia, and heart rate following oxygen administration in healthy young adults. *Physiol. Behav.* **67**, 783–789 (1999).
40. Chung, S.-C. *et al.* Effect of 30% oxygen administration on verbal cognitive performance, blood oxygen saturation and heart rate. *Appl. Psychophysiol. Biofeedback* **31**, 281–293 (2006).
41. Tsunoda, K., Chiba, A., Yoshida, K., Watanabe, T. & Mizuno, O. Predicting changes in cognitive performance using heart rate variability. *IEICE Trans. Inf. Syst.* **E100-D**, 2411–2419 (2017).
42. Mezzacappa, E. S., Kelsey, R. M., Katkin, E. S. & Sloan, R. P. Vagal rebound and recovery from psychological stress. *Psychosom. Med.* **63**(4), 650–657 (2001).
43. Taelman, J., Vandeput, S., Vlemincx, E., Spaepen, A. & Van Huffel, S. Instantaneous changes in heart rate regulation due to mental load in simulated office work. *Eur. J. Appl. Physiol.* **111**(7), 1497–1505. https://doi.org/10.1007/s00421-010-1776-0 (2011).
44. Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). North-Holland.
45. Eggemeier, F. T., and G. F. Wilson. Workload assessment in multi-task environments. In *Multiple task performance* (DL Damos, London, GB, 1991), pp. 207–216.
46. Gilbert, S., Bird, G., Frith, C. & Burgess, P. Does 'task difficulty' explain 'task-induced deactivation?'. *Front. Psychol.* **3**, 2012 (2012).
47. Lakens, D., Scheel, A. M. & Isager, P. M. Equivalence testing for psychological research: A tutorial. *Adv. Methods Pract. Psychol. Sci.* **1**, 259–269 (2018).
48. Taelman, J., Vandeput, S., Vlemincx, E., Spaepen, A. & Van Huffel, S. Instantaneous changes in heart rate regulation due to mental load in simulated office work. *Eur. J. Appl. Physiol.* **111**, 1497–1505 (2011).
49. Fikenzer, S. *et al.* Effects of surgical and FFP2/N95 face masks on cardiopulmonary exercise capacity. *Clin. Res. Cardiol.* **2020**, 1–9 (2020).
50. Hopkins, S. R., Stickland, M. K., Schoene, R. B., Swenson, E. R. & Luks, A. M. Effects of surgical and FFP2/N95 face masks on cardiopulmonary exercise capacity: The numbers do not add up. *Clin. Res. Cardiol.* **2020**, 1–2 (2020).
51. Kampert, M., Singh, T., Finet, J. E. & Van Iterson, E. H. Impact of wearing a facial covering on aerobic exercise capacity in the COVID-19 era: Is it more than a feeling?. *Clin. Res. Cardiol.* **2020**, 1–2 (2020).
52. Baig, A. S., Knapp, C., Eagan, A. E. & Radonovich, L. J. Jr. Health care workers' views about respirator use and features that should be included in the next generation of respirators. *Am. J. Infect. Control* **38**(1), 18–25 (2010).
53. Guo, Y. P. *et al.* Evaluation on masks with exhaust valves and with exhaust holes from physiological and subjective responses. *J. Physiol. Anthropol.* **27**(2), 93–102 (2008).
54. Grinshpun, S. A. *et al.* Performance of an N95 filtering facepiece particulate respirator and a surgical mask during human breathing: Two pathways for particle penetration. *J. Occup. Environ. Hyg.* **6**, 593–603 (2009).
55. van der Sande, M., Teunis, P. & Sabel, R. Professional and Home-Made Face Masks Reduce Exposure to Respiratory Infections among the General Population. *PLoS ONE* **3**, e2618 (2008).
56. Person, E., Lemercier, C., Royer, A. & Reychler, G. Effect of a surgical mask on six minute walking distance. *Rev. Mal. Respir.* **35**, 264–268 (2018).
57. Davies, H. J., Williams, I., Peters, N. S. & Mandic, D. P. In-Ear SpO2: A tool for wearable, unobtrusive monitoring of core blood oxygen saturation. *Sensors* **20**, 4879 (2020).
58. Stone, R. T. & Wei, C.-S. Exploring the linkage between facial expression and mental workload for arithmetic tasks. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* **55**(1), 616–619. https://doi.org/10.1177/1071181311551126 (2011).

59. Nourbakhsh, N., Wang, Y. & Chen, F. GSR and blink features for cognitive load classification. In *Human-Computer Interaction – INTERACT 2013*, 2013, pp. 159–166.
60. Marquart, G. & Winter, J. D. Workload assessment for mental arithmetic tasks using the task-evoked pupillary response. *PeerJ Comput. Sci.* **2015**(8), 16. https://doi.org/10.7717/peerj-cs.16 (2015).
61. Banbury, S. & Berry, D. C. Disruption of office-related tasks by speech and office noise. *Br. J. Psychol.* **89**(3), 499–517. https://doi.org/10.1111/j.2044-8295.1998.tb02699.x (1998).
62. Thomas, H. B. G. Communication theory and the constellation hypothesis of calculation. *Q. J. Exp. Psychol.* **15**, 173–191 (1963).
63. Chow, S.-C., Shao, J., Wang, H. & Lokhnygina, Y. *Sample Size Calculations in Clinical Research* 2nd edn. (CRC Press, 2017).

Acknowledgements
We thank Prof. Dr. Sebastian Möller for his support in making this investigation possible.

Author contributions
R.S. and K.P. created the test design, conducted the experiment, and analyzed the data. R.S. and K.P. wrote the main manuscript text, prepared all figures, and reviewed the manuscript.

Funding
Open Access funding enabled and organized by Projekt DEAL. Funding was provided by the Quality and Usability Lab, Technische Universität Berlin (Lab Head: Prof. Dr. Sebastian Möller). The Lab had no role in design and conduct of the study; collection, management, analysis, and interpretation of the data; and preparation, review, or approval of the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to R.P.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.