The Tiny Impact of Respiratory Masks - on Physiological, Subjective and Behavioral Measures under Mental Load

Robert Spang (Spang@tu-berlin.de)
Technical University of Berlin
Kerstin Pieper
Technical University of Berlin

Research Article

Keywords: COVID-19, face mask, viral spread

Posted Date: December 10th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-119883/v1

Abstract

Since the outbreak of the COVID-19 pandemic, the use of face masks is recommended to prevent droplets from traveling from one person to another. A commandment to wear masks applies in most public places to contain the widespread of the virus. Some public debates concern myths regarding risks caused by wearing a mask, 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 FFP2 (N95) respirator and wearing no mask. We focused on a demanding situation to show that cognitive efficacy and the person's state is equivalent in both conditions of a randomized-controlled cross-over study design. 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 broad as possible. We analyzed data from 44 participants 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 a mask and not wearing a mask.

Introduction

Through the COVID-19 pandemic, most countries quickly adopted – amongst others – face masks as a measure to protect the general public. The masks' major impact is to prevent droplets emitted by a person to travel close to a close-by second person [1]. By redirecting the exhaled by the mask, it is accomplished that aerosols better diffuse around one's head [2]. This reduces the threat of an infection (if other measures such as a sufficient distance are practices as well)[3],[4].

A significant part of societies all over the world question scientific findings, oppose government regulations, and spread misinformation regarding the coronavirus disease 2019 (Covid-19), its transmission, the consequence of infection, and numbers or associated deaths or intensive-care medicine capacities [5–8]. An early Twitter analysis estimated that around 25% of all tweets regarding Covid-19 contain misinformation [9].

Findings show that acquiring information about the pandemic on social media seem to result in a higher susceptibility to misinformation [10]. A study investigating the share of misinformation (related to Covid-19) in social media in the US showed that people tend to share misinformation due to failing to question the truthfulness of the content [11]. Besides the fact that this potentially promotes the spread of the virus, it also leads to extreme beliefs and conflicts. Especially the aversion to masks takes on dramatic dimensions and leads to senseless violence (https://edition.cnn.com/2020/05/06/us/security-guard-shot-mask-wife/index.html). This extreme attitude is supported by striking articles in the media like, e.g. (http://www.insidefortlauderdale.com/2233/Masks-Dont-Work) which uses specific results of a study by MacIntyre et al., 2015 [12] out of context and derives general statements like “Masks don’t work.” The actual aim of the study was to compare cloth masks to medical masks. Among other stated limitations of the study there was no comparison with a control group without masks and thus no evidence for the

allegedly linked statement. Such inflammatory contributions and the associated anti-mask attitude are not a phenomenon of our time; protest against masks also happened around 100 years ago during the influenza epidemic. The so-called "Anti-Mask-League" protested against the force to wear a mask as there were no sufficient scientific basis of their efficacy against the virus [13].

Therefore, our study aims to provide clarity and evidence against known myths. The scientific background at present shows that NIOSH filtering facepiece respirator (N95) / Filtering Face Piece (FFP2) / KN95 masks 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 a mask, but also for the people next to them [14]. Modeling the potential for wearing face masks clearly demonstrated a drastic decrease in peak hospitalizations and deaths, decreasing the virus's effective transmission rate [15].

A subjective evaluation of surgeons reported a hampered performance and decreased surgical fatigue while wearing FFP masks [16]. However, it remains unclear whether these observations are due to physiological (e.g. less oxygen supply) or psychological (e.g. reduced ability to communicate) factors. A prior study by Beder et al., 2008 [17] found a decrease in the blood oxygen saturation and an increase in pulse rates in surgeons wearing a mask pre- and post-surgery. However, as there was no control group (without mask) the findings are limited as they fail to discriminate if the investigated differences are caused by wearing the face mask or by stress. Another study that compared wearing a mask (N95) to not wearing a mask while exercising, 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 [18]. Solely end-tidal carbon dioxide (EtCO2) levels were increased while wearing a mask. The authors conclude that wearing a mask provides a feasible and safe way to exercise.

Given these findings, we hypothesize equivalence of blood oxygen saturation while wearing a mask and while not wearing one. Further, we hypothesize equivalence of the cognitive demand of the mask and the no-mask condition. Therefore, we expect that the participants perform equal in both test conditions. Thus, in terms of behavioral data we hypothesize equivalence between the conditions regarding: the number 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.

Findings from Scholey et al, 1999 [19] suggest that in the state of high cognitive demand the heart rate helps to regulate the metabolism the aim to increase circulation of blood oxygen and thereby improve cognitive performance. They showed that oxygen saturation and cognitive performance correlate with each other. Similar findings were made by Chung et al., 2006 [20], where hyperoxic air administration led to increased blood oxygen saturation and an improved accuracy in a verbal cognition task compared to normal air administration. In a different study the heart rate variability (HRV) was shown to be sensitive for different levels of cognitive performance and a higher HRV amplitude is suggested to contribute to a decrease in cognitive performance [21]. The task to be performed has a cognitive focus, and 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. Given the assumption that cognitive

demand is equal and that masks do not impact blood oxygen saturation, we hypothesize equivalent findings regarding participants' HRV.

Apart from the investigation of effects of facial masks, we contribute an exploratory analysis of the Apple Watch Series 6' (APW) blood oxygen level measurement function. We compare the readings of the APW with the CMS 50D, a clinical pulse oximeter approved by the American FDA for hospital use. The background of this comparison is to get an impression of how well a consumer product performs compared to a clinical device. As wearable devices enjoy greater popularity, they become available for many private households. If the results are comparable, a recommendation could be derived to motivate private persons to explore different their blood oxygen values on their own. This confrontation could potentially positively affect the spread of false statements such as "mask makes breathing more difficult.".

We address the stated concerns using a randomized, crossover study design. We compare physiological, subjective and behavioral measures obtained while wearing an FFP2 (N95) mask and without. During both conditions, participants engage in a demanding, cognitive task: solving mental arithmetic problems while being pressed for time. The effectiveness of inducing mental load is measured using a short form of the NASA Task Load Index (NASA-TLX) and compared against a baseline measurement.

Results

For all following Two One-Sided Test of Equivalence (TOST) procedures, we employed equivalence boundaries of $d_z = \pm .45$. 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 confidence intervals of the TOST and the Null hypotheses significance tests, together with the equivalence boundaries.

For both HRV analyses, we had to exclude 3 datasets due to incomplete recordings. Hence, both are based on data from 41 participants. Given the chosen alpha level of $\alpha = .05$ and the pre-defined equivalence bounds of $d_z = \pm .45$, both HRV TOST results have a statistical power of $1 - \beta = .784$. All other test are based on data from all 44 participants, resulting in a statistical power of the TOSTs of $1 - \beta = .820$.

Physiological data

See Fig. 2 for a visualization of the blood oxygen saturation and the two HRV metrics per condition.

Physiological: Blood Oxygen Levels

The mean difference of blood oxygen level between wearing a mask and not to do so immediately after performing the 15 min of mental calculation is − .3% (Median: 0%, SD: 1.64%). The increase of blood oxygen level without a mask has a negligible effect size of $d_z = -.119$. A Shapiro-Wilk test indicated a

violation of the assumption of normality ($W = .919$, $p = .004$), 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 $d_z = \pm .45$, which translates to $\pm .736$ in the units of the metric at hand (percent in this case). Hence, the lower equivalence boundary $\Delta_L = -.736\%$ and the upper equivalence boundary $\Delta_U = .736\%$. The TOST procedure reveals that the effect observed is statistically equivalent, the larger of the two p values is less than $\alpha = .05$ ($V = 682$, $p = .014$). According to the Neyman-Pearson approach, this means that one can reject the hypothesis that the true effect is greater than $d_z = \pm .45$ and act as if the effect size falls within these equivalence bounds [22]. Accordingly, to our pre-registration, we additionally run an exploratory null hypothesis significance test. A pairwise Wilcoxon signed rank test returned nonsignificant ($V = 154$, $p = .259$), hence the H0 of no difference between groups is not rejected.

Physiological: heart rate variability (RMSSD & SI)

The mean difference of RMSSD between wearing a mask and not wearing one immediately after the 15 min of mental calculation is 6.73 ms (Median: 1.63 ms, SD: 35.69 ms). The decrease of the RMSSD without a mask has a negligible effect size of $d_z = .153$. A Shapiro-Wilk test indicated a violation of the assumption of normality ($W = .375$, $p < .001$), 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 $\alpha = .05$ ($V = 56$, $p < .001$). We additionally ran an exploratory null hypothesis significance test. A pairwise Wilcoxon signed rank test returned nonsignificant ($V = 519$, $p = .257$).

The mean difference of Baevsky's Stress Index (SI) between wearing a mask and not wearing one immediately after the 15 min of mental calculation is 17.28 (Median: 2.39, SD: 67.26). The decrease of the SI in conditions without a mask has a negligible effect size of $d_z = .108$. A Shapiro-Wilk test indicated a violation of the assumption of normality ($W = .819$, $p < .001$), hence we employed a robust TOST procedure using Wilcoxon signed rank test, with equivalence bounds of $\Delta_L = -30.265$ and $\Delta_U = 30.265$. It reveals that the effect observed is statistically equivalent, the larger of the two p values is less than $\alpha = .05$ ($V = 237$, $p = .005$). We additionally ran an exploratory null hypothesis significance test. A pairwise Wilcoxon signed rank test returned nonsignificant ($V = 523$, $p = .240$).

Figure 2: Comparison of the physiological metrics (blood oxygen level and HRV) while wearing a mask and not wearing a mask. 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.

Exploratory Analysis: accuracy of the Apple Watch Series 6 blood oxygen level readings

Out of the 180 measurement time points (four times per participant), the Apple Watch was unable to determine the blood oxygen levels 10 times (5.68% of the cases). We repeated each measurement until

we obtained a value, up to five times at most. Of these failed 10 cases, each of the five consecutive measurements resulted in an 'Unsuccessful Measurement' error message.

Out of the remaining, successful readings, we had to conduct 1.523 measurements on average (median: 1.0, SD: 0.955) to obtain a successful reading. However, 69.16% of the successful measurements succeeded at the first go, the remaining 30.84% of the records needed at least one repetition.

The difference between the Oximeter and the Apple Watch readings indicate a small bias towards overestimated blood oxygen values of the Apple Watch (Mean: -0.639%,

Median: -0.5%, SD: 3.093). Because a Shapiro-Wilk tests rejected the assumption of normality for the 166 difference records (statistic: 0.877, $p < .001$, skew: -0.844, kurtosis: 5.504), we performed a Wilcoxon signed rank test. It indicated a significant bias of the Apple Watch for higher blood oxygen readings ($W = 2869.5$, $p = .004$). The inner 95% of the differences of the two devices span a range of 11% (from − 7–4%). This is reflected in the Bland-Altman plot, Fig. 3. Differences seem to decrease as the average increases. This indicates a greater disagreement for lower blood oxygen readings.

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 − .01 (Median: .08, SD: 1.74, see Fig. 4). The decrease of the TLX score without a mask has a negligible effect size of $d_z = -.002$. The assumption of a normal distribution was not rejected ($W = .988$, $p = .919$), so we used a TOST procedure based around Welch's paired t-test with equivalence bounds of $\Delta_L = -.782$ and $\Delta_U = .782$. It reveals that the effect observed is statistically equivalent, the larger of the two p values is less than $\alpha = .05$ ($t(43) = 2.956$, $p = .003$). An exploratory null hypothesis significance test (pairwise Welch's t-test) returned nonsignificant ($t(43) = -.029$, $p = .977$).

Figure 4: Comparison of the subjective load ratings (NASA TLX) while wearing a mask and not wearing a mask. While the distribution reveals minor differences between the groups, these are averaged out when comparing mean and median descriptives.

Behavioral data

See Fig. 5 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 a mask against while not wearing one is -2.14 (Median: 3.5, SD: 19.75). The increase of correct responses in conditions without a mask has a negligible effect size of $d_z = -.075$. The assumption of a normal distribution was rejected (Spahiro-Wilk test, $W = .935$, $p = .015$), so we used a robust TOST procedure based around the Wilcoxon signed rank test with equivalence bounds of $\Delta_L = -8.887$ and $\Delta_U = 8.887$. It reveals that the effect

observed is statistically equivalent ($V = 680$, $p = .016$). An exploratory null hypothesis significance test (pairwise Wilcoxon signed rank test) returned nonsignificant ($V = 496$, $p = .995$).

Behavioral: ratio correct responses / all tasks

Now, we investigate the ratio of correct responses against the number of all responses given (correct and incorrect). The mean difference between mask and no mask is nearly zero (Median: − .01, SD: .08). The effect induced by the mask is negligible ($d_z = .028$). The assumption of a normal distribution was not rejected ($W = .96$, $p = .133$), so we used a TOST procedure based around Welch's paired t-test with equivalence bounds of $\Delta_L = -.035$ and $\Delta_U = .035$. It reveals that the effect observed is statistically equivalent, the larger of the two p values is less than $\alpha = .05$ ($t(43) = -2.635$, $p = .005$). An exploratory null hypothesis significance test (pairwise Welch's t-test) returned nonsignificant ($t(43) = 0.349$, $p = .728$).

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 mask and no mask is nearly zero (Median: − .01, SD .1). The effect induced by the mask is negligible ($d_z = -.005$). The assumption of a normal distribution was not rejected ($W = .977$, $p = .524$), so we used a TOST procedure based around Welch's paired t-test with equivalence bounds of $\Delta_L = -. 043$ and $\Delta_U = .043$. It reveals that the effect observed is statistically equivalent, the larger of the two p values is less than $\alpha = .05$ ($t(43) = 2.922$, $p = .003$). An exploratory null hypothesis significance test (pairwise Welch's t-test) returned nonsignificant ($t(43) = -0.063$, $p = .950$).

Behavioral: mean response time

The mean difference between mask and no mask of the average response time is .29 s (Median: − .05 s, SD: 1.39 s). The decrease of the response time in conditions without a mask has a small effect size of $d_z = .214$. The assumption of a normal distribution was rejected (Spahiro-Wilk test, $W = .911$, $p = .002$), so we used a robust TOST procedure based around the Wilcoxon signed rank test with equivalence bounds of $\Delta_L = -.626\ s$ and $\Delta_U = .626\ s$. It reveals that the effect observed is statistically equivalent ($V = 329$, $p = .026$). An exploratory null hypothesis significance test (pairwise Wilcoxon signed rank test) returned nonsignificant ($V = 529$, $p = .699$).

Behavioral: mean response time of correct responses

Lastly, we investigate the average response time of only correct responses. The mean difference between mask and no mask is .2 s (Median: − .03 s, SD: 1.12 s). The decrease of the response time in conditions without a mask has a negligible effect size of $d_z = .183$. The assumption of a normal distribution was rejected (Spahiro-Wilk test, $W = .929$, $p = .009$), so we used a robust TOST procedure based around Welch's paired t-test with equivalence bounds of $\Delta_L = -.506\ s$ and $\Delta_U = .506\ s$. It reveals that the effect observed is statistically equivalent ($t(43) = -1.831$, $p = .037$). An exploratory null hypothesis significance test (pairwise Welch's t-test) returned nonsignificant ($t(43) = 1.154$, $p = .255$).

Discussion

The blood oxygen saturation shows a slight decrease of .3% after wearing an FFP2 mask. This effect is statistically insignificant. Although some discussions against the use of facial masks argue that masks would impair the body's oxygen supply, this is unstrained by our findings. Instead, we found statistical equivalence and no difference between the mask conditions. The HRV metrics, RMSSD and as Baevsky's SI (sometimes known as Index of Regulation Strain) showed statistical equivalence when comparing the mask against the no-mask condition, as well as no significant difference from each other. When focusing on the descriptive values, both HRV metrics seem to decrease slightly (statistically insignificant) in the no-mask condition.

When interpreting the HRV metrics as indicators of mental load, the RMSSD typically drops if the participant is more strained, while the SI typically increases under the same condition [23],[24]. Given our results, we find opposing interpretations: while the RMSSD indicates more strain, higher intensity load, and focus in the no-mask condition, the SI indicates the opposite, being reduced in numeric value and indicating a more relaxed state of mind in the same condition. This paradox interpretation underlines that the changes induced by the facial mask cause less variability than the HRV metrics can interpret reasonably.

The subjective NASA-TLX ratings show that the participants perceived a statistically equivalent workload between wearing a mask 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 a mask and not. We did not explicitly tell them about our research question before the experiment was over. However, some participants might have figured out the reason why to wear a mask sometimes and why not (none of the participants implied so). But 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 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., [25] reported "marked discomfort" of the participants wearing FFP masks, although the study, in general, is heavily debated, see [26],[27]). Therefore, it seems to be an even greater confirmation that wearing a mask does not limit the wearer's performance. We deem it unlikely to confound all of our nine different metrics regarding the possible condition awareness, especially since we investigated a broad spectrum of varying measurement dimensions.

Regarding the behavioral data the mask's influence reached a small effect ($d_z = .214$) 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 mask 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 minutes when not wearing a mask nor was their ratio of correct responses any better. Even the average response time was statistically equivalent to the mask 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 a mask in one condition, we decided to provide them with CE-certified FFP2 masks (N95 in the US / KN95 in China) for the mask condition. FFP2 masks generally sit tighter on the face, suffer from less face seal leakage, and its filter medium offers stronger filter characteristics than surgical masks [28]. Since most homemade masks have even less powerful filtering properties than surgical masks [29], two interpretations can be drawn for wearers of these simpler masks: Either the FFP2 mask itself mostly attributes the effects that we 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 facial masks.

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 [18],[30] 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 $d_z = .5$). 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 pre-registration before assessing the recorded data and even before most of the participants had been sampled (as recommended by [22]). However, a replication with smaller equivalence boundaries and larger sample size would further substantiate our findings.

Other than that, it's worth pointing out that our mental load condition lasted only 15 minutes. In discussions with mask-skeptic people we heard the argument that wearing masks for a whole day would impact cognitive functioning. Clearly, our comparison based around two 15 minutes 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. [31]. Hence, if there is no evidence for any impact of the masks after wearing them for 15 minutes, there is little reason to believe that this drastically changes after several hours.

Finally, we compared the blood oxygen level readings of the Apple Watch Series 6 with a clinical oximeter. This device (CMS 50D) provides an accuracy description of ± 2%, compared to a gold standard measurement (blood gas analysis). When comparing the wrist-worn wearable to the finger oximeter, this imperfection of the reference device has to be taken into account. However, the limits of agreement between the two devices with 12,12% are much larger than the accuracy of the clinical device. Given these wide bounds, the results are ambiguous, and the devices are not essentially equivalent. The increase in differences with a decreasing average indicates a ceiling effect since the typical blood oxygen saturation of healthy adults is around 95%, and the readings can't exceed the 100% mark. Hence, more

variability below 95% is to be expected. This explains in parts why the variability isn't consistent across the graph. To conclude, a precise blood gas analysis is needed to assess the device's accuracy thoroughly. However, the number of repetitions necessary to get reliable results, as well as the fail-rate and the comparison to our reference device, suggest that the Apple Watch's blood oxygen level is more a vague estimation than a reliable clinical tool.

We hypothesized that wearing a mask while performing a demanding, cognitive task for 15 minutes does not statistically differ from completing the same task without a mask. To do so, we created a test bed 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 $d_z = .45$. We interpreted that we can reject the hypothesis of a large effect induced by a mask (larger than $d_z = .45$). 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 a mask and not wearing one. The combination of physiological, subjective, and behavioral data delivers a measurement tool that allows us to detect potential differences both 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.

Methods

Task

The cognitive task consisted of solving basic arithmetic equations (addition, subtraction, multiplication, or division) presented visually on a display in front of the participants. The response time was limited, depending on the estimated difficulty of the task. This estimation was done using a prediction model based around the Q-value of Thomas, 1963 [32] to estimate the difficulty of a basic arithmetic task. This task design allowed us to induce a constant, high mental load for the duration of each condition.

Procedure

We experimented in a small and bright lab room in Technical University of Berlin. The current pandemic situation forced us to limit the amount of time spent with the participants to < 15 minutes. Hence, the general introduction was done in a separate room and the time together was spend to instruct the conditions. Before and after each experiment, the lab room was heavily ventilated and all surfaces and devices have been 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) and a smartwatch (model: Apple Watch Series 6; Apple Inc.). 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 mask we provided were 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).

The stimulus presentation was done via smartphone (model: iPhone XR; Apple Inc.) with a 6 × 1-inch screen. To keep the contact between participant and experimenter to a minimum, we instructed the participant on using the Apple Watch blood oxygen measurement app by themselves. As this had to be done after each test condition, the instruction enabled us to avoid additional non-necessary contacts. The measurement on the watch takes 15 seconds, but the pulse oximeter is continuous. Therefore, we instructed the participants to note down the result of the Apple Watch and the current reading of the oximeter at the very moment the Apple Watch measurement finishes. In case of an erroneous reading, they had to repeat it up to 5 times.

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 like 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. 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. Apart from the baseline and the two measurements after each condition, we performed an additional, fourth blood oxygen measurement to have more than 150 paired records of the Apple Watch and the pulse oximeter. This added recording was solely done to have a substantial number of readings for the accuracy assessment. To compare the impact of the mask, we only used the readings obtained immediately after each condition.

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

Twenty-four of the 45 participants identified themselves as female. The mean age was 30.3 years (Median: 29.0, SD: 9.432, ranging from 20 to 64). 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 years, 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 was evaluated in a fast-track process and classified as harmless 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; informed consent was obtained from all participants before the recordings began.

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 a mask 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 preregistration. 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 (Lakens, Scheel & Isager, 2018). In other words, the equivalence test is used to examine whether the difference between wearing a mask and not wearing one is at least as extreme as a mid-sized effect of $d_z = \pm .45$.

We defined the SESOI as $d_z = .45$, based on analyzing reported effect sizes of the related literature (especially [18],[30]). However, previous studies had a slightly different focus, which is why 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 of $d_z = 0.5$. 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).

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

Declarations

Acknowledgements

We thank the members of our chair for the support, inspiring discussions, and continuous encouragement to pursue our ideals.

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.

Competing Interests

The authors declare no competing interests.

References

[1] Leung, N. H. L. *et al.* Respiratory virus shedding in exhaled breath and efficacy of face masks. *Nat. Med.* **26**, 676–680 (2020).

[2] Asadi, S. *et al.* Efficacy of masks and face coverings in controlling outward aerosol particle emission from expiratory activities. *Sci. Rep.* **10**, 15665 (2020).

[3] 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).

[4] Morawska, L. & Milton, D. K. It is time to address airborne transmission of COVID-19. *Clin Infect Dis* **6**, ciaa939 (2020).

[5] 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).

[6] Mian, A. & Khan, S. Coronavirus: the spread of misinformation. *BMC Med.* **18**, 1–2 (2020).

[7] Sharma, K. *et al.* Coronavirus on social media: Analyzing misinformation in Twitter conversations. *ArXiv Prepr. ArXiv200312309* (2020).

[8] Tasnim, S., Hossain, M. M. & Mazumder, H. Impact of rumors or misinformation on coronavirus disease (COVID-19) in social media. (2020).

[9] Kouzy, R. *et al.* Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus* **12**, (2020).

[10] Roozenbeek, J. *et al.* Susceptibility to misinformation about COVID-19 around the world. *R. Soc. Open Sci.* **7**, 201199.

[11] 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).

[12] MacIntyre, C. R. *et al.* A cluster randomised trial of cloth masks compared with medical masks in healthcare workers. *BMJ Open* **5**, e006577 (2015).

[13] Dolan, B. Unmasking History: Who Was Behind the Anti-Mask League Protests During the 1918 Influenza Epidemic in San Francisco? *Perspect. Med. Humanit.* **5**, (2020).

[14] 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).

[15] 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).

[16] 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).

[17] 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).

[18] 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).

[19] 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).

[20] 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).

[21] 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).

[22] Lakens, D., Scheel, A. M. & Isager, P. M. Equivalence testing for psychological research: A tutorial. *Adv. Methods Pract. Psychol. Sci.* **1**, 259–269 (2018).

[23] 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).

[24] Baevsky, R. M. & Chernikova, A. G. Heart rate variability analysis: physiological foundations and main methods. *Cardiometry* (2017).

[25] Fikenzer, S. *et al.* Effects of surgical and FFP2/N95 face masks on cardiopulmonary exercise capacity. *Clin. Res. Cardiol.* 1–9 (2020).

[26] 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.* 1–2 (2020).

[27] 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.* 1–2 (2020).

[28] 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).

[29] 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).

[30] 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).

[31] 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).

[32] Thomas, H. B. G. Communication theory and the constellation hypothesis of calculation. *Quarterly Journal of Experimental Psychology* **15**, 173–191 (1963).

Figures

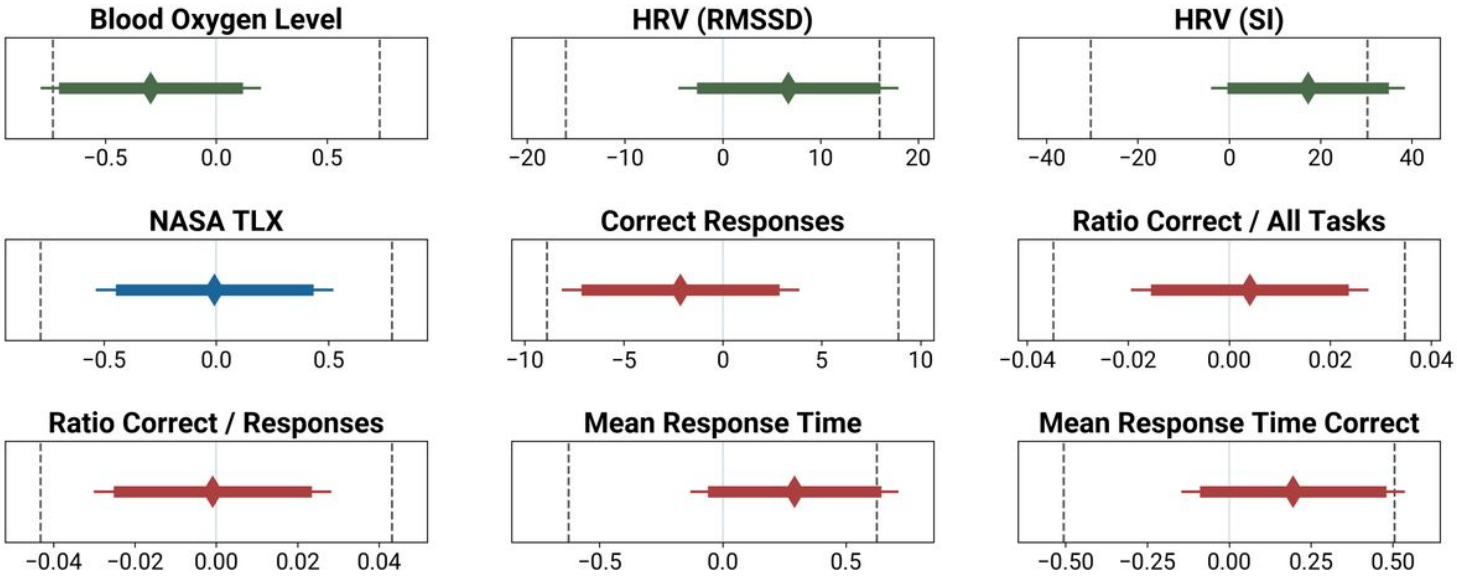

Figure 1

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.

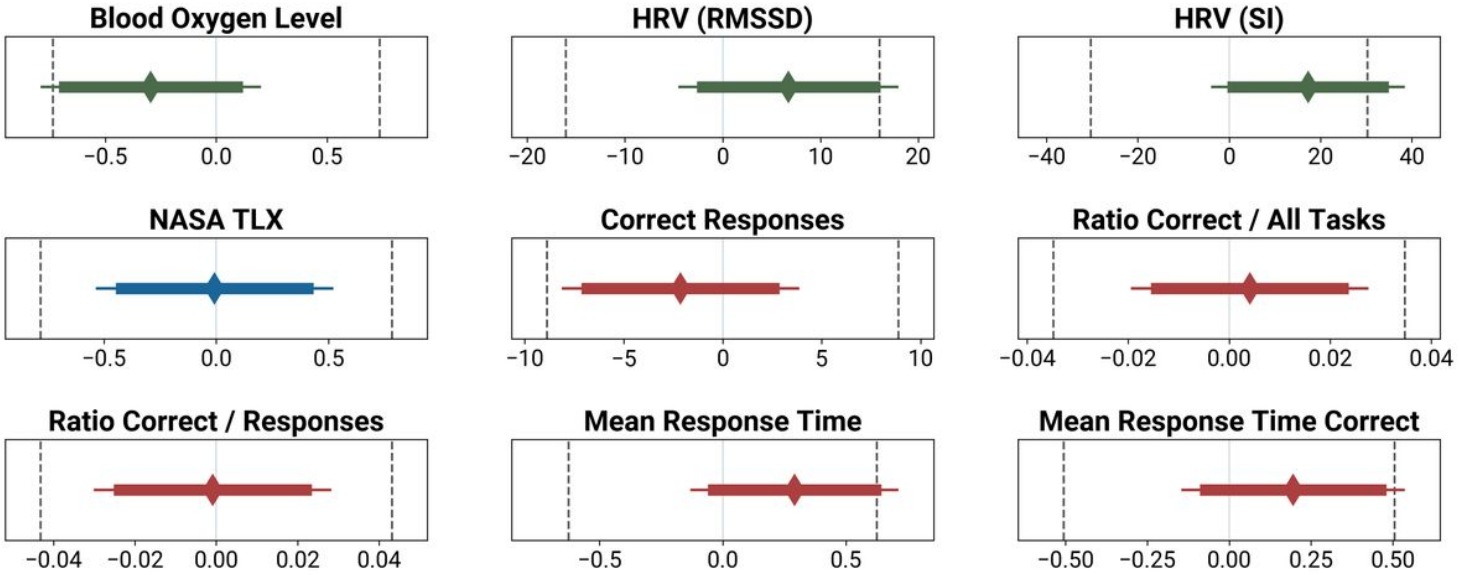

Figure 1

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.

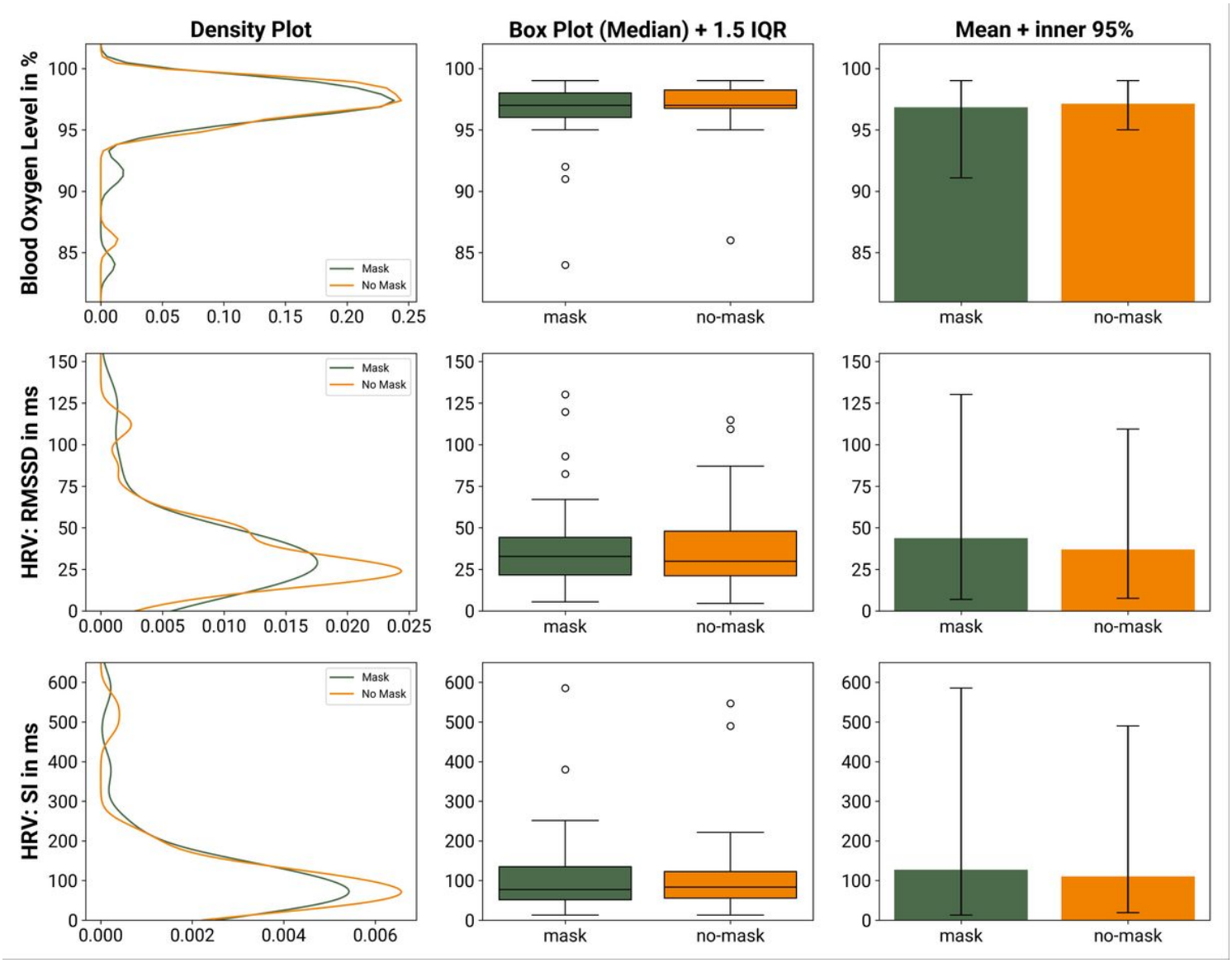

Figure 2

Comparison of the physiological metrics (blood oxygen level and HRV) while wearing a mask and not wearing a mask. 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.

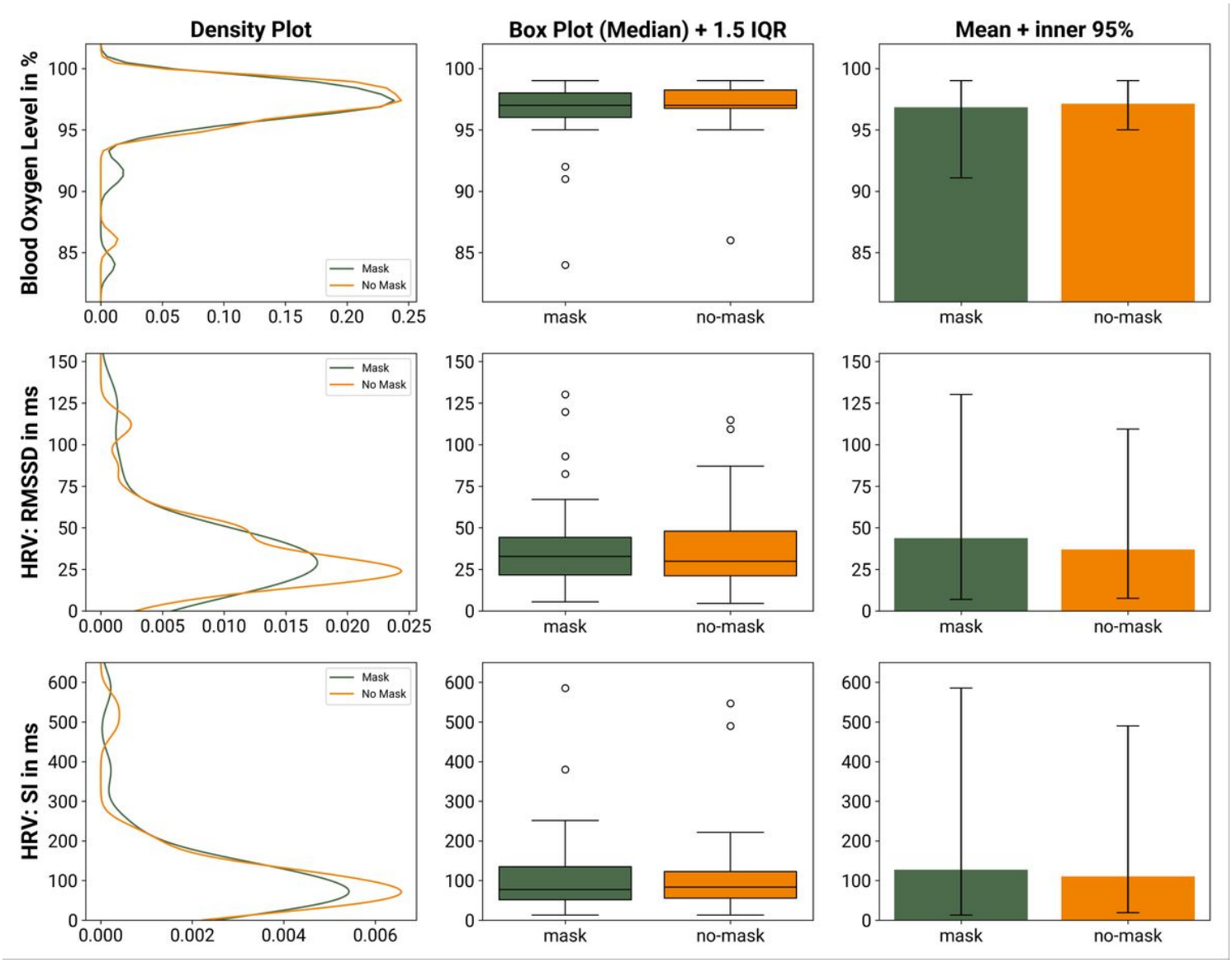

Figure 2

Comparison of the physiological metrics (blood oxygen level and HRV) while wearing a mask and not wearing a mask. 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.

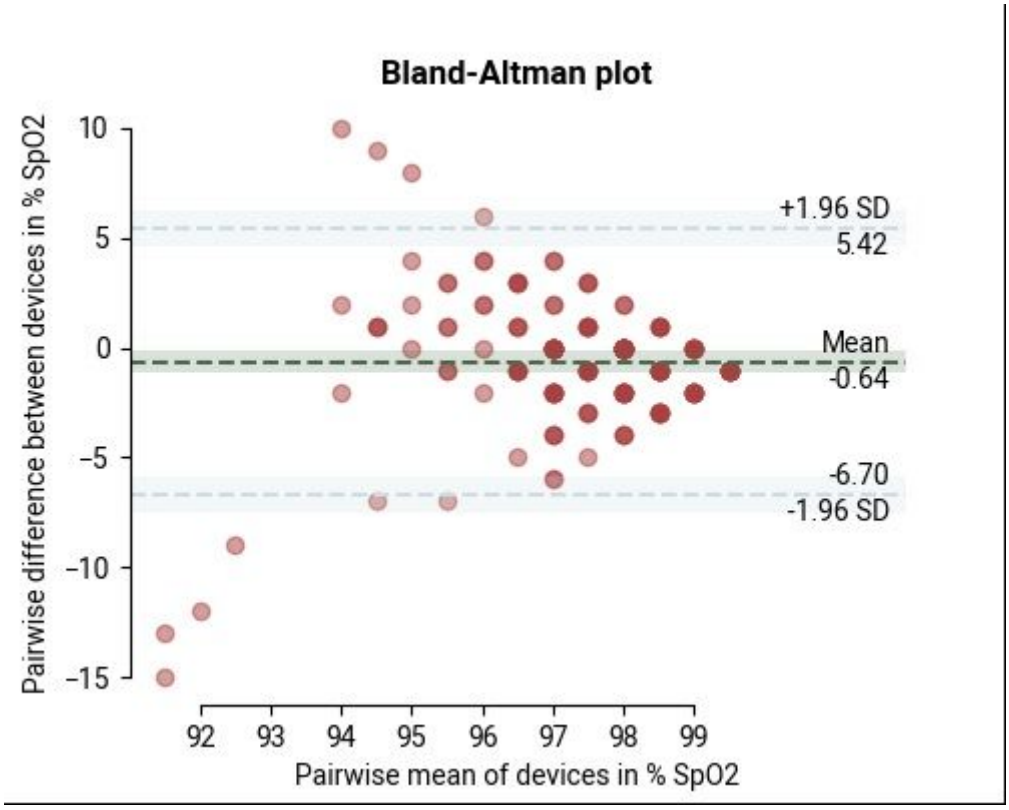

Figure 3

The Bland-Altman plot compares the mean of two devices' readings with their difference. The angelfish-shaped assembly of the records indicates increasing reading differences with decreasing mean blood oxygen levels. Note, both devices compared only provide integer readings. This is why many of the dots form a raster.

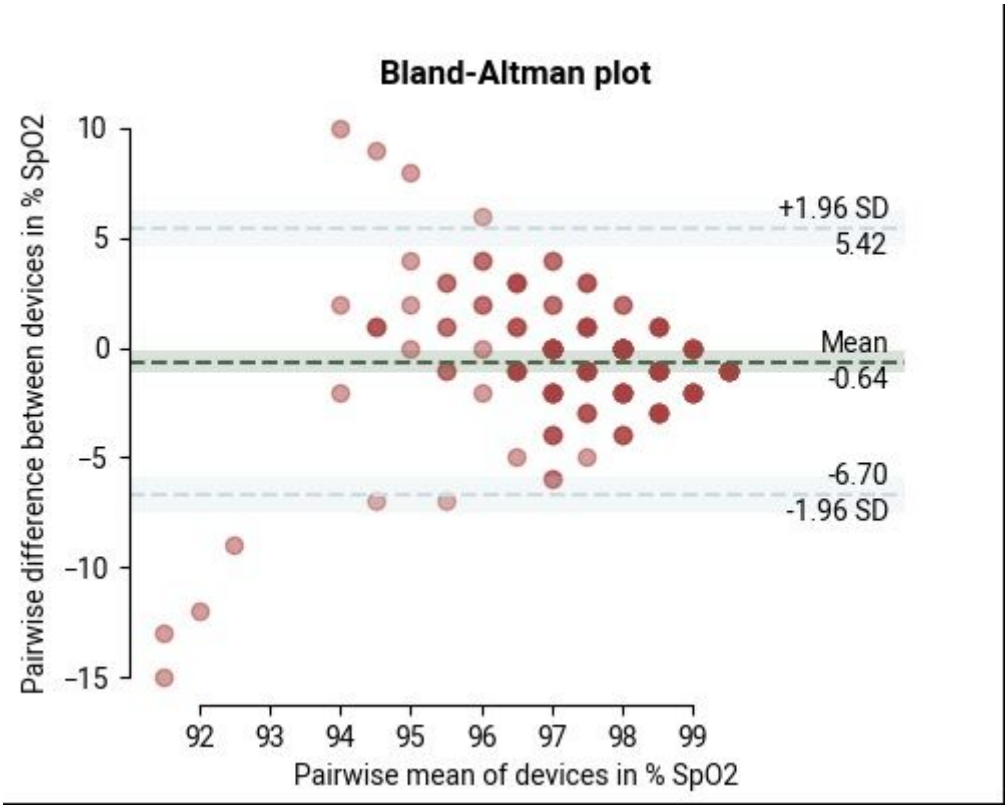

Figure 3

The Bland-Altman plot compares the mean of two devices' readings with their difference. The angelfish-shaped assembly of the records indicates increasing reading differences with decreasing mean blood oxygen levels. Note, both devices compared only provide integer readings. This is why many of the dots form a raster.

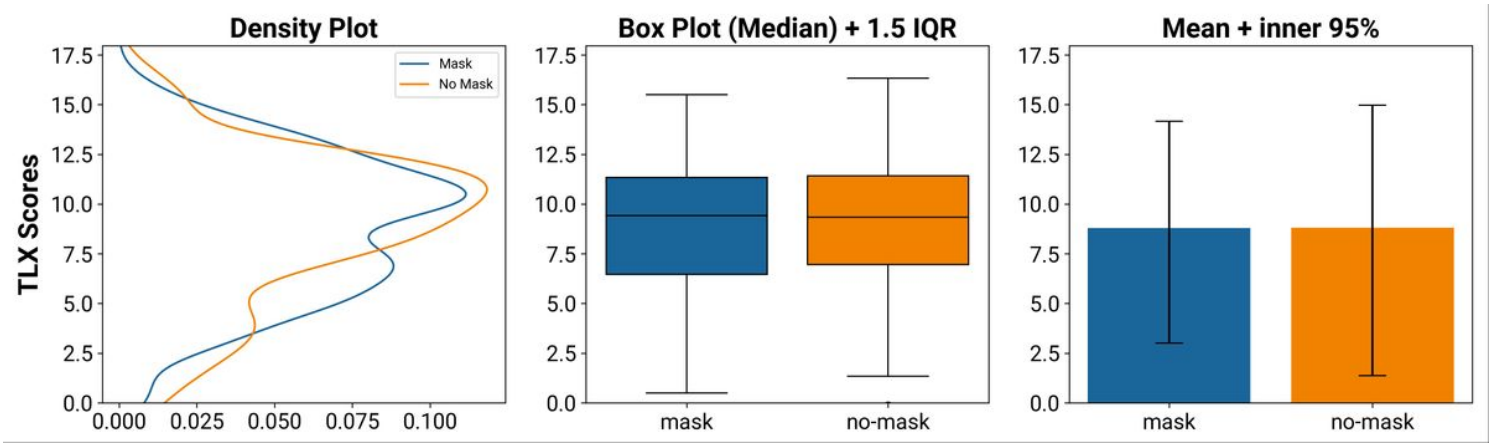

Figure 4

Comparison of the subjective load ratings (NASA TLX) while wearing a mask and not wearing a mask. While the distribution reveals minor differences between the groups, these are averaged out when comparing mean and median descriptives.

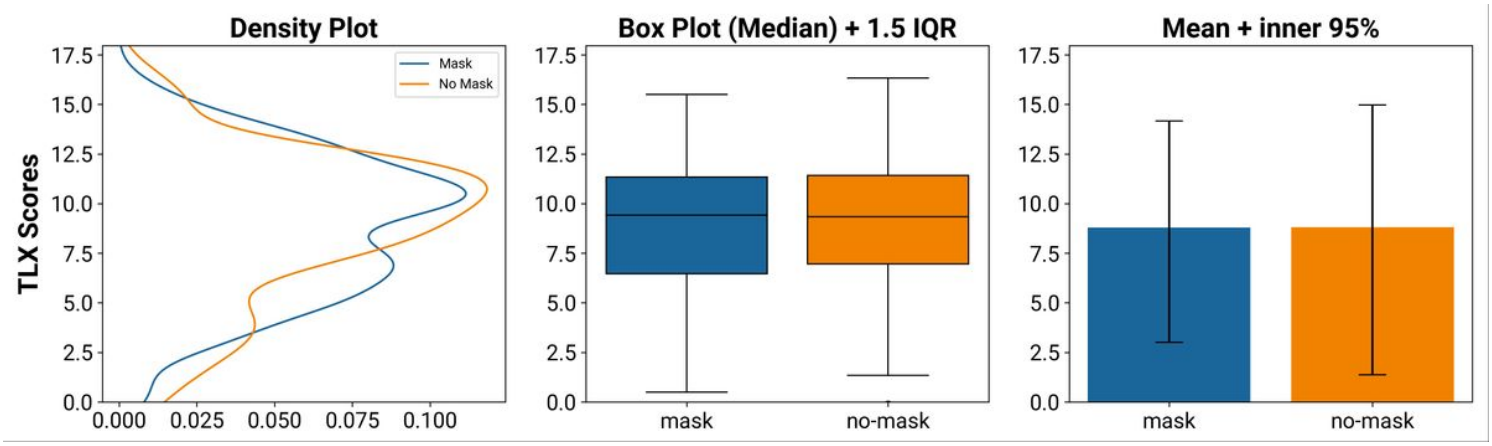

Figure 4

Comparison of the subjective load ratings (NASA TLX) while wearing a mask and not wearing a mask. While the distribution reveals minor differences between the groups, these are averaged out when comparing mean and median descriptives.

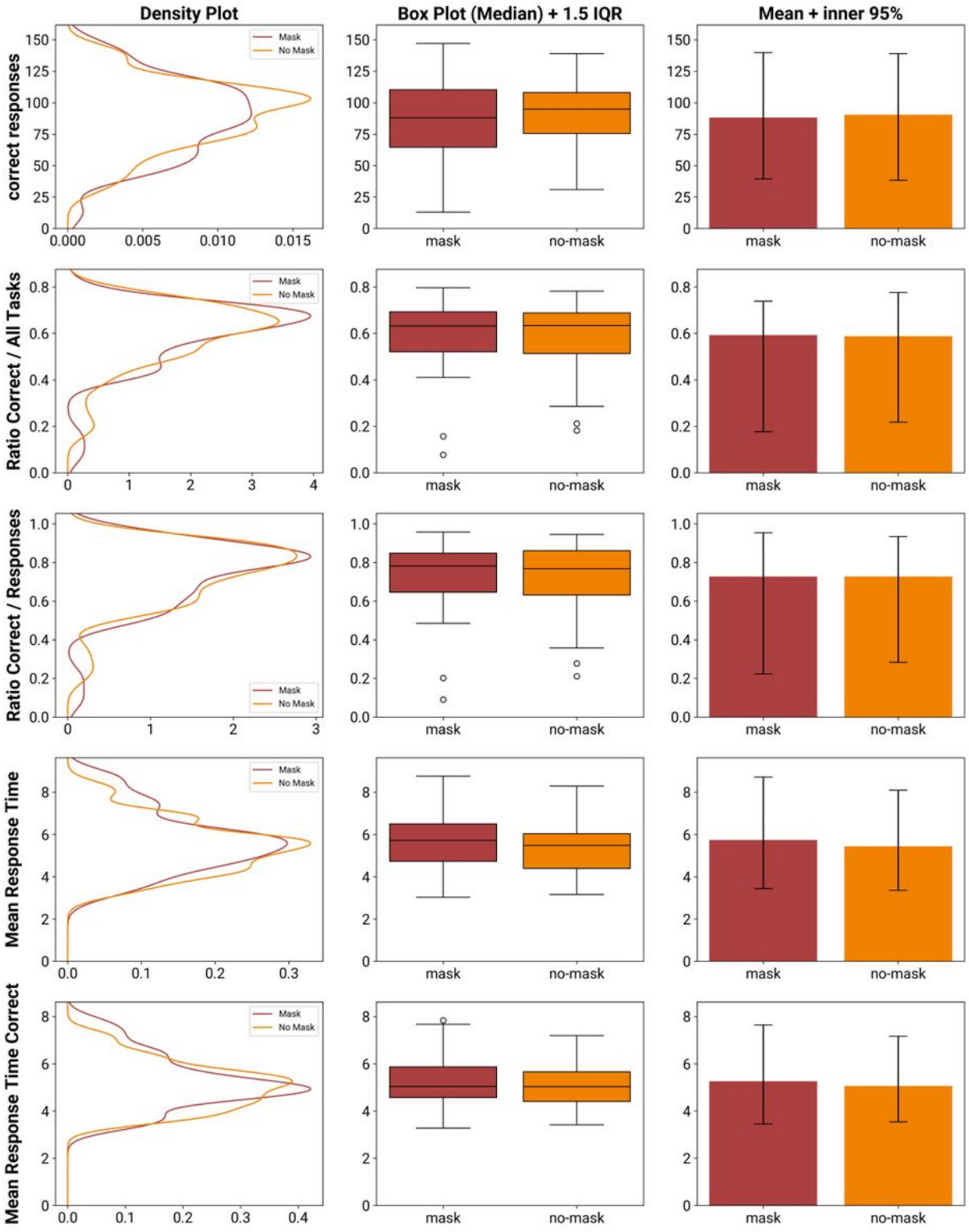

Figure 5

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.

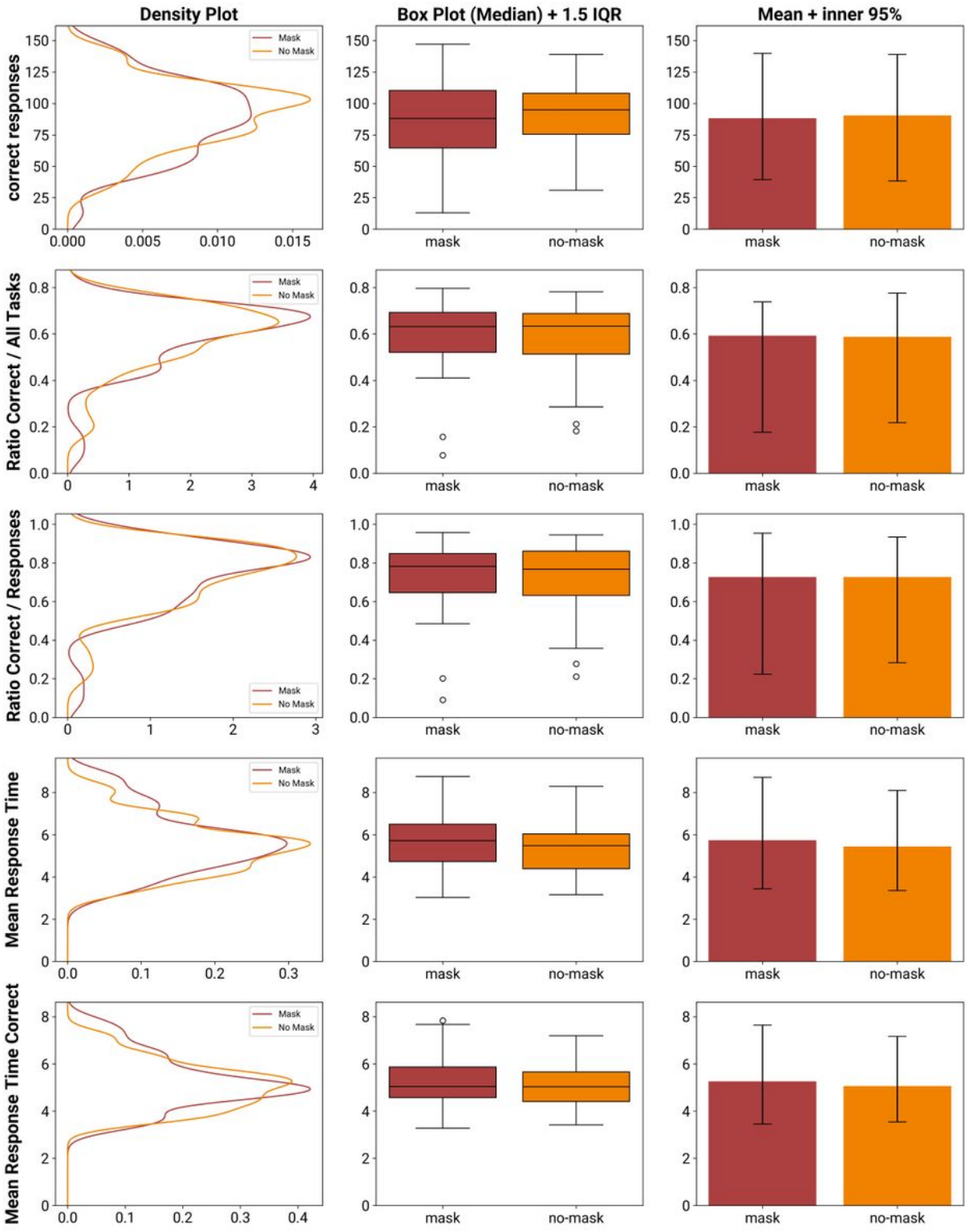

Figure 5

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.