Turning Up the Heat: The Impact of Indoor Temperature on Cognitive Processes and the Validity of Self-Report

Martijn Stroom*[†] Nils Kok[†] Martin Strobel[‡] Piet M. A. Eichholtz[†]

Abstract

Indoor climate interventions are often motivated from a worker comfort and productivity perspective. However, the relationship between indoor climate and human performance remains unclear. We assess the effect of indoor climate factors on human performance, focusing on the effects of indoor temperature on decision processes. Specifically, we expect heat to negatively influence higher cognitive rational processes, forcing people to rely more on intuitive shortcuts. In a laboratory setting, participants (N=257) were exposed to a controlled physical environment with either a hot temperature (28ºC) or a neutral temperature (22ºC), in which a battery of validated tests were conducted. We find that heat exposure did not lead to a difference in decision quality. We did find evidence for a strong gender difference in self-report, such that only men expect that high temperature leads to a significant decline in performance, which does in fact not materialize. These results cast doubt on the validity of self-report as a proxy for performance under different indoor climate conditions.

Keywords: indoor climate, heat, performance, decision quality, heuristics, biases, risk-taking, self-report

1 Introduction

Performance at work is influenced by many factors such as individual characteristics, leadership, experienced work pressure, incentive schemes, and corporate structure (Hermalin & Weisbach, 1991; Perry & Porter, 1982; Wageman & Baker, 1997). The physical climate of the workplace is typically not mentioned as an important factor. This is remarkable, as office buildings have been undergoing rigorous innovations throughout recent decades (for instance, see Vermeulen & Hovens, 2006). Developments in insulation, ventilation, and air-conditioning quality are effectively changing the indoor environment to which workers are exposed. These innovations are typically motivated from a building efficiency and/or worker comfort perspective, but while there is ample research highlighting the impact of increased energy efficiency on building resource consumption (Eichholtz, Holtermans, Kok, 2019; Pérez-Lombard, Ortiz, & Pout, 2008), the link between changes in indoor environmental conditions and human performance remains a topic of debate (MacNaughton et al., 2017; Satish et al., 2012; Zhang et al., 2017).

Research regarding the impact of indoor environment on worker performance is hampered by the fact that high-skilled performance measures at work are difficult to obtain directly, and are hard to compare between disciplines. For example, Zivin & Neidell (2012) show that pear-pickers' performance suffers from exposure to bad environmental quality conditions. However, the output of highly skilled workers who face cognitively demanding tasks – such as academics, managers, doctors, or investors – lacks such direct outcome measure. Moreover, any output that is measurable is not easily traced back to a quantifiable time period of exposure to the physical indoor climate. It is exactly this type of high-skilled workers that spends considerable time in confined offices or meeting rooms, subject to specific indoor climate conditions.

To circumvent the challenge to correctly assess human performance, research has shifted from measuring performance to comfort (Bluyssen, 2013). The implicit expectation is that when the climate is rated as "comfortable", productivity increases. Comfort measures are an attractive proxy for productivity and performance, as they are easily and inexpensively assessed by self-report. However, whether self-assessed comfort levels are indeed an accurate proxy for performance remains an open question. Psychological re-

The authors would like to thank seminar participants at Maastricht University, Juan Palacios, Wouter van Marken Lichtenbelt, Rick Kramer, the members of the 2019 Social Judgement and Decision Making Conference, as well as participants at the doctoral poster session of the American Real Estate and Urban Economics Association during the 2020 Allied Social Sciences Association Conference. We are especially grateful for the valuable comments of Caroline Goukens. We thank Sustainably.io for the use of their indoor climate sensor 'Birdnest'. This paper is financially supported by Maastricht University's Graduate School of Business and Economics Primary Data Collection Seeding Grant 2019.

*Corresponding Author, Email: m.stroom@maastrichtuniversity.nl, ORCID iD: 0000-0003-3411-4260

[†] School of Business and Economics, Department of Finance, Maastricht University.

[‡] School of Business and Economics, Department of Economics, Maastricht University.

search repeatedly suggests self-reported introspection into one's own subjective experience and emotions to be unreliable (Engelbert & Carruthers, 2010).

In this paper, we assess the effect of indoor environmental conditions on human performance, by investigating decision processes. Tversky and Kahneman (1974), amongst others, distinguish decision making as "intuitive" and "rational" processes. Automated, intuitive rules of thumb, or heuristics, are "quick and dirty" and applied without much effort. The rational processes need more time and cognitive resources, are only scarcely applied, and are also associated with high decisional quality. A mainstream application of the interplay between these fast and rational or effortful processes is the default-interventionist approach (Evans, 2007). It stipulates that the effortful processes can intervene in the fast heuristics, when a wrongful application in a given context is detected (also known as bias). Thus, whenever the effortful processes are hampered, for instance due to cognitive constraint resulting from environmental factors, increased bias-susceptibility generally lowers overall decisional quality (Gawronski & Bodenhausen, 2006; Muraven & Baumeister, 2000).

1.1 Literature

1.1.1 Temperature and Cognition

Psychological and neurological research has attempted to identify the effects of temperature on cognitive functions. We elaborate on two relevant findings.

The most profound and general finding is that cognitive capacity is lowered by adverse temperature conditions. Wright, Hull, & Czeisler (2002) find that changes in the temperature of the body and brain are correlated with changes in performance, such that deviating temperatures from the internal optimal will worsen performance. Shibasaki, Namba, Oshiro, Kakigi, & Nakata (2017) show that neurological inhibition processes suffer from heat stress. In decision-making, executive and inhibition processes coordinate which stimuli to act on (execute) and which not (inhibit). Both these biological processes are found to be less strong under heat stress. Van Ooijen, Van Marken Lichtenbelt, Van Steenhoven, & Westerterp (2004) suggest that temperature could influence mental performance as a result of fatigue. This view is similar to the theoretical concept of mental depletion, the cognitive model stipulating limited mental "control" resources for self-regulation (Baumeister Bratslavsky, Muraven, & Tice, 1998). Mental depletion often results in more instinctive behaviour (such as aggression; Van Lange, Rindery, & Bushman, 2017). In this context, adverse temperature conditions could drain cognitive capacity due to the mental effort needed to compensate for the adverse context. In general, when external stimuli overstimulate, concentration and performance become more costly (MacLeod, 1991). Indeed, Cheema & Patrick (2012) show that temperature generally lowers cognitive performance, but not for people that were already mentally depleted at the start of the task. Although mental depletion is debated (Carter, Kofler, Forster, & Mccullough, 2015; Hagger et al., 2016), the general notion of negative cognitive performance effects after enduring strain on mental capacity seems to be widely accepted (Cunningham & Baumeister, 2016).

The second key finding of research on temperature and cognition is that not all mental processes are affected equally. Lowered cognitive capacity appears theoretically very close to behavioural fatigue. However, it is important to understand that these two concepts are fundamentally and hierarchically distinct. When discussing behavioural fatigue, we consider a general lowering of behavioural activity (i.e. a 'global' effect). Decrease of cognitive capacity does not have a general uniform effect, but is depending on the neurological area that suffers most (i.e. a ' local' effect). Lan, Lian, Pan, & Ye (2009) find performance to decrease with adverse temperatures, but the effects differ across tasks.

In sum, it is clear that temperature has a general, or global, effect on cognition and cognitive performance, and that some local effects can be identified as well.

1.1.2 Temperature and Intuition

The literature review by Hancock & Vasmatzidis (2003) suggests that high capacity and complex mental processes are more profoundly affected by temperature than automated processes. Automated tasks rely on a strong and fast relation between stimulus and response, making them less susceptible to mental constraints (Kahneman, 1973). Automated tasks are part of system I in Kahneman's cognitive framework – also known as the intuitive system. They rely on intuition and on simple rules of thumb that are learned and are often successfully applied to predictable situations. System II is slow and costly on mental resources, but is generally associated with high-quality decision making.

Cognitive capacity and cognitive control are highly correlated (Engle & Kane, 2003), and the latter has also been found to be affected by temperature. Shibasaki, Namba, Oshiro, Kakigi, & Nakata (2017) show that neurological inhibition processes suffer from heat stress. In decision making, inhibition and executive processes coordinate to achieve an optimal solution. As such, the effect of heat on performance can be twofold: not only do higher-order complex tasks suffer more than simple automated tasks (Grether, 1973), but wrongful application of an automated process or application of a wrong automated process might be less likely to be corrected. In other words, even when the direct effect of heat on simple and automated processes is not evident (as stated by Zhang & de Dear, 2017), the outcome can still suffer in quality due to the lack of high order process intervention. Indeed, Hancock & Vasmatzidis (1998) find that highly skilled operators suffer less from performance decrease under heat stress, and they argue that this is most likely a result of performance depending on automated internalized processes.

The cognitive framework of Tversky and Kahneman leads to relevant predictions when we apply the findings of temperature on tasks complexity and intuition. The interaction found between temperature and automated tasks and task complexity suggests that system I could be less affected than system II. The default-interventionist approach (Evans, 2007) states that both system work parallel to each other, and system II generally attempts to identify mistakes made by system I and intervenes if necessary. Consequently, the wrongful application of heuristics increases, because the controlling function of system II would fail as the system suffers from temperature.

We therefore expect that the distinct effect that heat has on cognition can be (partially) captured by the Kahneman framework. Recent research has investigated the effect on cognitive reflection (Chang & Kajackaite, 2019), but to date, no study has extended this investigation to the behavioural biases stemming from a predisposition to overly adhere to intuitive decision strategies. To our knowledge, no attempts have been made to distinguish the effects of heat on behaviour and cognition using this approach.

1.1.3 Temperature and Gender

The effect of temperature on cognition is heterogeneous for gender. Biological research (Kingma & Van Marken Lichtenbelt, 2015), as well as metabolic research (Byrne, Hills, Hunter, Weinsier, & Schutz, 2005) and psychological empirical research (Wyon, 1974) shows that hot temperatures have a distinctly different effect on women as compared to men. The most profound example of this distinction and its neglect in the past decade is the temperature comfort level. The 'default' room temperature level of 21ºC seems mainly based on male preferences (Kingma & Van Marken Lichtenbelt, 2015). Indeed, anecdotal evidence suggests that women perform better at slightly higher default room temperatures (Chang & Kajackaite, 2019).

As such, finding the effects of adverse temperature on cognition would be incomplete without taking gender-specific preferences into consideration. Without correcting for gender, female preference or tolerance for higher temperatures might influence the overall findings regarding the effect of adverse temperatures on performance. Given that women show a preference for somewhat higher temperatures, women will rate identical absolute temperature increases (subjectively) as less adverse as compared to men. Performance for women might thus also be expected to be less affected by heat.

1.1.4 Temperature and Risk

Evidence suggests that temperature has a direct effect on the willingness to take risk. Wang (2017) shows that people making trading decisions will pursue high-risk high-yield options compared to a control condition.

Some indirect evidence on aggression also suggests that risky behaviour could follow from loss of control through the same channel. For instance, solely increasing the temperature makes people subjectively rate other people in the room to be more hostile (Anderson, Dorr, DeNeve, & Flanagan, 2000). Cao & Wei (2005) hypothesize that aggression leads to increased risk behaviour. Denson, DeWall, & Finkel (2012) conclude that it is the loss of self-control that increases aggression. Finally, Frey, Pedroni, Mata, Rieskamp, & Hertwig (2017) show self-control to be predictive of various risk behaviour outcomes. Overall, we expect the same channel that increases system I dependency will also increase risk-taking behaviour.

1.2 This study

We hypothesize that heat exposure will decrease human 'performance' such that biased behaviour will be more prominent, as rational correction will require more effort under heat stress. Heat is a salient factor in the working environment and workers can often elicit control over temperature themselves, making the relevance of our results apparent and immediately applicable. Moreover, by testing detectable differences in conditions, we are able to assess the relevance of self-reported comfort measures.

Additionally, we investigate the effect of heat on risk behavior. Through the same channel, we expect that a combination of lack of effortful control and bodily discomfort will increase risk behaviour. This would be in line with aggression studies (for instance, American football players commit more aggressive fouls; Craig, Overbeek, Condon, & Rinaldo, 2016). We test both the general self-reported risk attitude, which should be unaffected by the heat, given that it has been reported to be a rather stable character trait (Dohmen et al., 2011), and actual risk behaviour which we expect to increase following indoor temperature manipulation.

1.3 Experimental design

We design a controlled experiment to measure the effect of heat on decision quality. Participants (N=257) are exposed to a controlled physical environment with either a hot temperature (28ºC) or a neutral temperature (22ºC), in which a battery of validated test are conducted. These include cognitive reflection tasks, a heuristics battery, lottery risk tasks, and self-reported risk preferences. Additionally, participants state their personal comfort levels and their subjective estimation as to what extent the environment influences their performance on the battery of tasks.

Our experimental design has several key advantages over current practices in the literature. First, we actively strive to control a variety of factors influencing the physical experience of the environment. That is, we pre-expose all participants to the temperature manipulation for a defined adjustment period of one hour before starting the tasks. All participants are wearing similar clothing provided specifically for the experiment. We further control for the outdoor temperature of the period before testing. Second, we keep all other indoor climate factors constant. For instance, we manipulate the temperature while keeping air ventilation levels unchanged. As a result, CO_2 levels, noise, lighting, and air refreshment are equal between manipulations. Some recent experiments manipulated temperature by opening and closing windows, without controlling for CO_2 and fine particles between groups, and are therefore unable to isolate the effect of just temperature on task performance (Wang, 2017).

2 Method

2.1 Experimental conditions and design

We employed a stratified random sampling method to recruit a total of 257 participants with an average age of 21.57 (SD = 2.41) years old using the Maastricht University Behavioral Experimental Economics laboratory database. Stratification ensures an equal gender distribution amongst manipulation groups. The final sample allows for a 10% deviation of gender within groups. Participants are randomly distributed to either the control or the experimental condition. This between-subject design uses temperature as the main independent variable. Given the clear gender differences in the temperature effect on performance and satisfaction in the literature, gender is the secondary independent variable in our analysis.

The experiment is programmed using Qualtrics Software (Qualtrics, Provo, UT) and executed at the Behavioral Experimental Economics lab facilities at Maastricht University, the Netherlands. The laboratory is approximately 5 meters wide and 20 meters long. In this room, there are 33 cubicles (approx. 1.0 meter by 1.5 meters), all including a computer and table, which are closed off by shutters. Air quality is controlled using a climate system that holds the air refreshment rate constant. The control condition of 22 ºC is reached running only the climate system. The "hot" condition of 28 ºC is reached using five 3 kW industrial heaters, each with a 115m^3 capacity. During the experiment, four heaters maintain a constant temperature. Manual adjustments to the thermostats of the individual heaters ensures a stable temperature. All heaters also ran without heating during the control condition, such that the noise produced by the heaters is constant between conditions.

All participants are subject to strict clothing prescriptions. These requirements ensure that all participants have a similar physical experience of the heat. For instance, the possibility to remove layers of clothing could increase heterogeneity in the experienced heat within and between conditions. All participants are asked to wear long jeans. To fully ensure homogeneity, we provide all participants with long-sleeved black polyester thermoshirts. Participants are not allowed to wear anything underneath these shirts.[1]

Furthermore, all participants arrive in the laboratory at 11 AM, one hour before the start of the actual experiment. This adaption time ensures that all participants experience the indoor climate similarly, independent of the outdoor temperature or previous activity. Moreover, the outdoor temperature is measured on all testing days and compared between conditions (see Appendix table 4 for an overview of the outdoor temperature between conditions). The tasks are given in the order in which they are presented in Section 2.2.

2.2 Dependent measures

2.2.1 Performance measures

Cognitive Reflection Task – The classic cognitive reflection task (CRT) by Frederick (2005) measures participants' propensity to rely on intuition or rational thinking. The test consists of three questions, of which each question has a salient intuitive answer and a correct rational answer. Each of these questions are scored binary, with 1 for a correct response or 0 for a biased and thus wrong response. The total score for this task is the total amount of correctly answered questions, such that the score of the CRT lies between 0 (no correct answers) and 3 (all answers correct). Although this test is often used, Bialek & Pennycook (2017) find that multiple exposure does not reduce its validity.

Cognitive Reflection Task Expansion – To increase the probability of capturing the distinction between intuitive and rational thinking in our sample, we add an expansion of the original CRT. This test (see Toplak, West, & Stanovich, 2014) consists of three additional items, following the same structure. It has been shown to be highly correlated to the original CRT.

Heuristics Battery – The heuristic bias task battery by Toplak, West, & Stanovich (2011) include various questions about well-known economical biases. We select ten questions from this battery concerning casual base rate neglect, sample size problems, sensitivity towards regression to the mean, framing bias, outcome bias, conjunction fallacy, probability matching, ratio bias, methodological reasoning, and the covariation problem.[2] Each of these questions are scored binary, with 1 for a correct response or 0 for a biased and thus wrong response. All scores of these questions are then added up. The resulting score on this battery is between 0

[1] Women are allowed to wear bras underneath.

[2] For an overview of these tasks, see Toplak et al., (2011)

and 10 points (M = 6.32, SD = 2.16), in line with the original authors.

2.2.2 Risk measures

Risk Elicitation Task – The first measure of risk assessment is aimed at inducing or eliciting actual risk behaviour at the time of the experiment. Similar to the original task of Holt & Laury (2002) we show the participants nine choices between two sets of lotteries. The first lottery is of relatively low risk, where both the high and low payout options diverge only minimally (€6 versus €4.80, respectively). The second lottery can be considered high risk, as there is a strong divergence between the high (€11.55) and low (€0.30) payout option. For each consecutive choice, the probability of the high payout in both lotteries increases with 10%, such that in the first choice the probability of the high payout for each lottery is 10% and in the ninth and final choice this probability has become 90%. Note that the expected payout of the high-risk lottery surpasses the payout of the low-risk lottery from step 5 onwards (since then the expected payout is €5.93 for the high-risk versus €5.40 for the low-risk lottery). Participants are scored on a scale from 1-10, where the score reflects the switching point of the participants. Score 1 indicates a sustained preference for the high-risk lottery, labelling them as "risk-loving". A score of 5 implies risk-neutral behaviour, as participants follow the switching point in which both measures are equivalent. A score of 10 is assigned when participants never switch to the high-risk lottery. We label these participants as "risk averse". Depending on the risk preference, all scores are considered rational, as even in step 1 or 9 there is still a 10% probability of a high win or loss, respectively. This lottery is incentivised, and participants are told that one of the lottery choices will be played at the end of the questionnaire. The outcome of their chosen lottery will be added to their total reimbursement. To make this incentive at least 25% of the total reimbursement, the lottery outcomes are multiplied by a factor from the original (Holt & Laury, 2002).

Risk Attitude Task – In addition to a risk elicitation task, we ask participants how risk-loving they perceive themselves to be, both in general and on specific domains. Participants rate themselves on a 10-point Likert scale, with the lowest score being risk-averse, and the highest score labelled fully prepared to take risk. First, all participants state to what extent they are willing to take risk or avoid taking risk generally as a person. Second, their willingness to take or avoid risk are specified for the following domains: driving, financial matters, leisure and sport, their occupation, health, and faith in other people. This approach has been extensively validated and proven to correlate with actual risk behaviour (Dohmen et al., 2011; Falk, Dohmen, & Huffman, 2016). Participants who switched their choice of lottery more than once are excluded from the sample, and 34 observations were thus excluded (16 male, 18 female).

2.2.3 Indoor climate satisfaction

Self-reported Indoor Climate Satisfaction and Hinder – Self-reported indoor environmental satisfaction is assessed by adapting the occupant indoor environment quality survey developed by Berkeley's Centre for the Built Environment (Huizenga, Abbaszadeh, Zagreus, & Arens, 2006). For temperature, air quality, noise, and lighting, all participants are asked to rate their satisfaction level on a Likert scale from 1 to 7. Additionally, for all these factors, participants are asked to what extent they perceive it as hindering or supporting their ability to answer the questions in the questionnaire on a similar 7 point Likert scale. The scores are recoded such that a score of 7 indicates that the factor fully supports their ability, and a score of 1 indicates that the factor fully hinders their ability to answer the questionnaire. We label the totality of these factor-specific measures "satisfaction measures". In the analysis, we control for multiple testing.[3]

2.2.4 Additional checks

CRT multiple exposure check. After the three performance tasks (e.g. original CRT, extended CRT, and the Heuristics battery), all participants are asked to indicate whether they recognize any if these questions and if yes, whether they also remember the correct answer. These questions are scored by 1 – yes, 2 – no, or 3 – unsure.

Clothing check. All participants are asked to indicate whether they are indeed wearing the thermoshirts provided by the experimenter.[4] On a Likert-scale of 1 (bad) to 7 (good), participants indicate the fit, length, and the comfortability of the shirt. Additionally, we ask to what extent the shirt influences the performance on the tasks using the same scale.

Temperature. To be able to check for climate adjustment effects, three questions aim to assess the current and past climate experienced by the participants as well as their climate preference. Specifically, participants are asked to state in which country they grew up (most time spend until your 18th birthday), in which country they lived for the majority of the last five years, and what their preferred thermostat setting is (in degrees Celsius) in winter.

[3] Multiple testing correction is applied for all 10 conditions using the Benjamini & Hochberg procedure (Benjamini & Hochberg, 1995), see Appendix table 7. This procedure aims to control the false discovery rate whilst preserving relatively higher power compared to more conservative procedures (e.g. Bonferroni correction; Thissen, Steinberg, & Kuan, 2002).

[4] One of the participants indicated to be allergic to the fabric of the thermoshirts, and was thus asked to wear a similar (long-sleeved) shirt. All other participants wore the thermoshirts provided by the experimenter.

2.3 Incentives payoff

The payout is determined by adding the outcome of the preferred lottery of the risk elicitation task to the standard endowment of €15. The participants are told that for one of the steps, their chosen lottery will be played, but do not know which step this will be. The Qualtrics Internal Randomizer is used to draw an outcome (50/50 allocation) for the lottery chosen by the participant at step 5. The outcome is displayed at the end of the questionnaire. For the whole sample the average expected payoff of the risk task is 27% of the total payoff (with mean €5.98). No other performance tasks are incentivised, as these specific tasks are found not to be affected by incentives (Brañas-Garza, Kujal, & Lenkei, 2019).

3 Results

3.1 Descriptives and Condition Manipulations

The recorded sample consists of 257 students ranging from 17 to 31 years old, of which 53.5% are female (see Appendix Table 3).[5] The recorded indoor and outdoor climate conditions are reported in Appendix Table 4. The average temperature in the control condition is 22.4°C and in the hot condition 28.3°C. Levels of indoor CO_2, outdoor temperature of each test day during the morning, and outdoor temperature of the past three days do not differ significantly between manipulations.

3.2 Satisfaction measures

We first present the climate satisfaction measures in Table 1. Looking at the first column, it is confirmed that temperature (d= 0.77) and air quality (d= 1.53) are assessed to be significantly less satisfactory in the hot condition. Additionally, both are predicted to hinder the performance on the performance measures. This confirms the notion that the high-temperature manipulation is considered uncomfortable.

Looking at the other indoor factors, and taking male and female participants together, we do not observe lighting satisfaction to be significantly different between conditions. The same holds for the effects of light on perceived performance. Similarly, we find no difference for noise satisfaction between conditions. However, it is reported to improve performance in the hot conditions. Here also, we note that noise was kept constant between conditions. Interestingly, participants actually predict noise to improve performance compared to the control condition. We suggest that in the control condition, when the heaters only produced noise, participants perceive the noise on its own as potentially hindering performance. In the hot conditions the noise of the heaters may be driven to the background by the more salient temperature. Also, in the hot condition there is a justification for the noise. Finally, we observe that clothing satisfaction and hinder do not differ between conditions.

3.3 Gender Differences and Temperature

Following recent studies of gender differences and temperature effects on performance, we examine the satisfaction measures when controlling for gender. Interestingly, the general dissatisfaction and increased hinder of temperature are reflected in our male sample only. These findings are presented in the middle two columns of Table 1. Our results are in line with Chang & Kajackaite (2019), such that males dislike hot temperatures and report to suffer more from heat as compared to women. This notion is further supported by the observation that temperature experience differs between genders when related factors do not. When we compare air quality satisfaction and its hinder between the two conditions, we find that both men and women dislike the hot temperature condition equally compared to the control condition. We note that additional (marginally) significant inconsistencies are seen for rating factors that are stable between conditions such as noise and light. Those discrepancies are correlated with the temperature manipulation (e.g. a potential demand effect; also see limitation section).

Summarizing, we find that, as expected from the manipulations, temperature significantly lowers satisfaction and the perceived performance on the task, but only for the male sample. As such, as the commonly used hypothesis regarding the link between comfort and productivity predicts, we expect to find a decrease in performance on the performance measures for men, but not for women.

3.4 Performance Measures

Panel A of Table 2 shows the non-parametric results for the performance measures. We find no significant difference between control and hot conditions on any of the three performance measurements for the full sample. Only for women do we find a marginally significant difference (T=-1.75, p=0.08; d=0.30) between the performance on the CRT original between the control condition (M=1.26, SD=1.09) and the hot condition (M=1.61, SD=1.24).[6] Note that performance is increasing rather than decreasing. We conclude from these first results that the temperature has no direct effect on performance for men and women on our performance measures. If anything, we find weak support in line with Chang & Kajackaite (2019), as women seem to improve

[5] The sample shows a average self-reported math proficiency of 63 on a scale from 0 to 100

[6] For post-hoc effect size sensitivity analysis, see appendix table 10

TABLE 1: Main Results of Indoor Variables

				Men			Women		
	Control	Hot	*p*-value	Control	Hot	*p*-value	Control	Hot	*p*-value
Self-reported Indoor Variables Satisfaction and Hinder									
Temperature Satisfaction	4.66 (1.57)	3.50 (1.45)	.00***	5.13 (1.53)	3.05 (1.29)	.00***	4.25 (1.49)	3.90 (1.49)	.16
Air Quality Satisfaction	5.35 (1.18)	3.54 (1.41)	.00***	5.32 (1.23)	3.38 (1.39)	.00***	5.38 (1.15)	3.67 (1.44)	.00*
Light Satisfaction	5.33 (1.46)	4.95 (1.64)	.07	5.50 (1.55)	5.57 (1.03)	.56	5.19 (1.39)	4.42 (1.88)	.02*
Noise Satisfaction	5.36 (1.43)	5.57 (1.42)	.18	5.42 (1.51)	5.58 (1.39)	.58	5.30 (1.36)	5.55 (1.46)	.18
Clothing Satisfaction	5.71 (1.36)	5.55 (1.27)	.14	5.62 (1.37)	5.08 (1.33)	.02*	5.80 (1.37)	5.96 (5.96)	.81
Temperature Hinder	4.68 (1.54)	3.40 (1.49)	.00***	5.27 (1.25)	3.05 (1.25)	.00***	4.17 (1.60)	3.71 (1.62)	.07
Air Quality Hinder	5.07 (1.23)	3.71 (1.45)	.00***	5.03 (1.25)	3.65 (1.23)	.00***	5.10 (1.22)	3.75 (1.63)	.00*
Light Hinder	5.02 (1.55)	4.95 (1.58)	.77	5.12 (1.57)	5.45 (1.23)	.37	4.94 (1.54)	4.52 (1.72)	.20
Noise Hinder	4.94 (1.69)	5.36 (1.59)	.04	5.00 (1.77)	5.22 (1.65)	.52	4.88 (1.64)	5.48 (1.53)	.03*
Clothing Hinder	3.68 (1.29)	3.74 (1.25)	.89	3.93 (1.23)	3.75 (1.19)	.17	3.46 (1.31)	3.74 (1.30)	.30
Observations	*129*	*128*		*60*	*59*		*69*	*69*	

Note: all scores are on 1-7 Likert scale, and all scores are recoded such that 1 is bad or low, and 7 is good or high. Significance levels are based on nonparametric analysis. Standard deviation are given in parentheses. * indicates p-value < .05, ** p-value <.01, and *** p-value <.001, after multiple testing correction

rather than decrease their performance on one of the three tasks in the hot temperature condition. [7]

3.5 Risk measures

Risk preference elicitation task. As expected from a strong body of research (for an overview, see Byrnes, Miller, & Schafer, 1999), a baseline difference in risk behaviour is observed when comparing the control conditions as can be seen in Table 2, panel B. Based on parametric independent sample t-tests, men (M = 5.70, SD = 1.85) are significantly more risk-taking as compared to women (M = 6.48, SD =1.57; t = -2.42, p < 0.05; *d*=0.45), in line with the literature.

For the risk elicitation measure, participants in general do

[7] The results do show a clear and significant difference in CRT performance between genders. These results are in line with earlier findings (Brañas-Garza et al., 2019; Zhang, Highhouse, & Rada, 2016) and are suggested to be a result of gender difference in either math proficiency (for the self-reported math proficiency per gender, see appendix table 3; Welsh, Burns, & Delfabbro, 2013) or math self-efficacy (Brañas-Garza et al., 2019))

Table 2: Main Results of Performance and Risk Measures

				Men			Women		
	Control	Hot	*p*-value	Control	Hot		Control	Hot	*p*-value
Panel A. Performance Measures									
CRT original *(scored 0-3)*	1.67 (1.61)	1.76 (1.56)	.49	2.13 (1.07)	1.95 (1.03)	.34	1.26 (1.09)	1.61 (1.24)	.08
CRT Extended *(scored 0-3)*	1.53 (1.09)	1.71 (1.07)	.21	1.85 (1.04)	2.03 (1.02)	.33	1.26 (1.07)	1.42 (1.03)	.37
Heuristics Battery *(scored 0-15)*	6.34 (2.22)	6.26 (2.11)	.86	7.33 (2.12)	6.83 (1.98)	.18	5.48 (1.93)	5.83 (2.13)	.32
Observations	*129*	*128*		*60*	*59*		*69*	*69*	
Panel B. Risk Behaviour Elicitation									
Risk Elicitation *(scored 1-10: 1 = extremely risk-loving, 10 = extremely risk averse)*	6.11 (1.74)	5.90 (1.99)	.45	5.70 (1.85)	6.29 (2.05)	.12	6.48 (1.57)	5.61 (1.89)	.01*
Observations	*111*	*113*		*53*	*51*		*58*	*62*	
Panel C. Self-reported Risk Attitude									
General Risk Attitude *(scored 1-10: 1 = risk-averse, 10 = fully prepared to take risk)*	5.77 (1.91)	5.43 (1.75)	.12	6.08 (1.80)	5.40 (1.77)	.03*	5.49 (2.00)	5.46 (1.74)	.97
Observations	*129*	*128*		*60*	*59*		*69*	*69*	

Note: For all panels except C, all significance levels are based on parametric analysis. For panel C, significance levels is based on nonparametric analysis. Standard deviation are given in parentheses. * indicates p-vale < .05, ** a p-value <.01, and *** a p-value <.001

not differ between conditions. However, when we look at the gender subsamples, the picture changes. First, although men do not differ significantly in risk preference between conditions, women are significantly more risk loving in the hot condition (M = 5.61, SD = 1.89) compared to the control condition (M = 6.48, SD = 1.57; t = 2.75, p < .01 ; *d*= 0.50). As such, for women the risk and heat hypothesis appears to be a valid prediction.

When comparing the risk preferences of women in the hot condition with the control condition of male risk preference, we observe that women do not only become more risk loving in a hot condition, but that their risk preference becomes equal to that of men in a normal control situation.

General risk attitude. For the general risk attitude question "Are you generally a person who is fully prepared to take risks or do you try to avoid taking risks?" (See Table 2, panel C), men report to be less prepared to take risk when asked in a hot condition (Mdn = 6.5) compared to the control condition (Mdn = 6; z=2.1, p < .05; *d*=0.38).[8] This is surprising, as we explicitly ask participants to reflect on their general risk attitude. This question has repeatedly shown to be stable over time and context independent, and as such, is supposed

[8] Note that the risk aversion scores are inverse for both measures: In the general attitude measurement, a low score equates risk aversion, whereas in the risk elicitation measure, a high score shows a late (or no) switch to the risky lottery, synonymous for risk averse behaviour according to the authors of the measure.

to be a stable predictor for risk behaviour. Women do report a stable attitude independent of conditions. [9]

When looking at the domain-specific risk attitudes, only one differs significantly between conditions: Men predict to be less risky on work-related issues in a hot condition (Mdn=6) compared to the control (Mdn=6.5; z =2.19 p=0.028; *d*=0.42) condition. For an overview of these results, see Appendix 5. This result remain significant when applying the Benjamini-Hochberg rank-dependent multiple testing correction (Benjamini & Hochberg, 1995) on the critical p-value threshold with a Q (false discovery rate) of 15%. [10]

4 Discussion

The increasing frequency of heatwaves, and outside temperatures that used to be exceptional, raises important questions about the impact of temperature on human performance. Of course, outdoor temperature does not need to be harmful given the mitigation effect of buildings, acting as a "shield" against temperature changes and pollution. There is evidence of a positive effect of building quality on human performance and productivity (for instance, see Palacios, Eichholtz, & Kok, 2020). But research measuring indoor climate also shows negative performance effects resulting from exposure to adverse indoor conditions (e.g. Künn, Palacios, & Pestel, 2019; X. Zhang, Wargocki, Lian, & Thyregod, 2017). Given that we spend roughly 90% of our time indoors, the effect of these adverse conditions warrants research. Understanding the effects of indoor temperature on human performance is crucial in determining and optimizing the daily indoor environment in work places and beyond.

The focus of this study is twofold: First, we assess the effect of hot temperatures on decision quality, and second, we answer the question whether peoples' stated experiences regarding these temperatures are related to this decision quality. In this study, we assessed the effect of adverse temperature by manipulation of the indoor temperature to 28ºC, compared to a control temperature of 22ºC.

From the expectation that rational decision-making would suffer under adverse temperatures, more reliance on intuition would lead to a lower score on the cognitive reflection task and to more biased responses in the heuristic battery. However, no significant difference on performance between the hot and control conditions are identified in this study. When looking at risk, a factor often associated with decisional quality and furthermore proposed to be correlated with the intuition-rational trade-off (Leith & Baumeister, 1996), we only observe an increase of risk preference in hot conditions for women.

Comparing these results with self-reported measures show some essential discrepancies. First, only men find the hot condition significantly less satisfactory as compared to the control condition. Women do not seem to make a distinction between conditions. Furthermore, when asking to what extent temperature has an influence on performance, men predict that the hot temperature significantly hinders their performance. Again, women do not make this distinction.

The discrepancy between self-report and actual behaviour is of crucial importance for the literature regarding the effects of indoor climate. Currently, self-reported measures are commonly used as a proxy for performance or productivity, yet this study shows that men are consistently overestimating the effect of adverse temperatures on performance. First, the discrepancy between the actual performance outcomes and the perceived hinder from adverse temperature for men shows that men would have expected to have performed better in the control condition, which they did not. If policy makers would have assessed this self-perceived hinder only, they might have spent significant effort and resources to improve indoor temperature conditions. In our study, however, we show that this would not result in an actual increase in performance.

On the domain of risk, we find that men assess their own daily willingness to take risk in general and in work situations to decrease when they are asked about this in the hot condition. This is surprising, since this measure is aimed at assessing the general self-reported risk preference, independent of any manipulation, and would thus be expected to be stable across conditions. For women, no significant difference between conditions is found. As for actual risk behavior, we find no difference between conditions for men.

These results have at least two implications for future indoor temperature (and indoor climate) research. First, we repeatedly find inconsistencies between the self-reported and actual effects of the indoor climate on performance. Specifically, men are overestimating the negative effect the temperature has on their performance. This shows that the use of self-reported measures as a proxy for actual performance is unreliable. Future research should focus on more direct measures of human performance and productivity than self-reported indoor climate satisfaction. Second, our research supports the recent findings of Chang & Kajackaite (2019) that gender plays an moderating part in the effect of temperature on performance. This underlines the conclusion from Kingma & Van Marken Lichtenbelt (2015) that one universal temperature standard does not fit the whole population. Gender differences have to be taken into account in any situation when we include temperature as an influential factor.

[9] When verifying the predictive power of the general risk attitude question with the risk behaviour as suggested by Falk, Dohmen, & Huffman, (2016), we find that in our sample the general risk attitude is not correlated with risk behaviour. Moreover, we find a negative correlation in the control condition between self-reported risk attitude and risk behaviour (see Appendix table 6). These result do not support the validity of the self-reported risk attitude as a proxy for risk behaviour.

[10] McDonald (2014) claims that a Q between 10% and 20% would entail relevant results, and underline that Q should not be mistaken for a P-value. For an overview of the critical value for 15% False Discovery Rate (Q) per rank used see Appendix 7.

4.1 Limitations

Three specific limitations are worth mentioning. First, our sample is restricted in size, background and age category. The sample size is limited as the adaption time required took more resources than in comparable studies. However, we are confident that addressing the exposure time is a key advantage of our experiment relative to the current literature. Regarding participant age, the sample mainly consists of students around the age of 22 (M = 21.57, SD = 2.41). We attempted to recruit an age category representing an older population (older than 50), but recruitment turned out to be difficult. Moreover, the level of English language skills and task comprehension forced us to exclude a significant part of the successfully recruited 'older' sample. The educational background of the majority of our sample (Business and Economics students) increased the likelihood of recognition of the type of tasks we assessed, and previous exposure to these constructs can influence results (we will discuss the results of multiple exposure to the CRT test below). Usage of the relatively unfamiliar extension of the CRT (Toplak et al., 2014) and a unfamiliar heuristic battery (Toplak et al., 2011) at least partially alleviates this concern.

Second, participants likely change behaviour in anticipation of the effect of the manipulation, which is unavoidable in an experiment with temperature manipulation. All participants in the manipulation conditions (e.g. the "hot" temperature condition), are instantly aware of this manipulation when entering the laboratory. To create uniformity between groups and take away emphasis on the temperature, we asked participants in all conditions to wear a provided shirt, and in both conditions the industrial heaters were on. Moreover, the indoor climate quality scale was not limited to temperature, but included other important indoor climate variables, reducing the emphasis on temperature. However, when the participants were asked to state what they thought the experiment was about, they indeed stated (in the manipulation condition) that temperature and task performance was the major aim of the experiment. In the control condition, less than 10% stated temperature to be a decisive factor (popular guesses included the influence of "clothing" or "noise" on performance).

Finally, the choice for our test battery is the outcome of a careful trade-off between practical and theoretical considerations. Research has suggested that the CRT is robust under multiple exposure (Bialek & Pennycook, 2017; Meyer, Zhou, & Frederick, 2018) and consistent over time (Stagnaro et al., 2018). Recognition of the original CRT is relatively high (46% recognized at least one question, and 20% recognized all questions) .[11] For the extended CRT questions, however, only 13% recognized one or more questions. The fact that we observe no difference in performance between the classic and extended CRT supports the notion that these levels of recognition and recollection of answers do not affect the results of this study.

Welsh et al. (2013) propose that the CRT merely reflects mathematical skills. In our sample we see that self-reported math skills differ significantly between genders. Women report a proficiency of 59.07 out of 100, whereas males report 67.48 out of 100 (p < .00). We indeed find that in the total sample, men outperform women in the CRT. However, this does not affect the result in the sense that we analyse the effect of temperature on performance specifically within gender. We furthermore find no interaction between math proficiency and the effect of temperature on the CRT. Nevertheless, we cannot exclude that the risk assessment is effected by the difference in math proficiency.

5 References

Anderson, C. A., Anderson, K. B., Dorr, N., DeNeve, K. M., & Flanagan, M. (2000). Temperature and Aggression. In Advances in Experimental Social Psychology (pp. 63–133). https://doi.org/10.1016/S0065-2601(00)80004-0

Baumeister, R., Bratslavsky, E., Muraven, M., & Tice, D. (1998). Ego Depletion: Is the active self a limited resource? Journal of Personality and Social Psychology, 74(3), 774–789. https://doi.org/10.1037/0022-3514.74.5.1252

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bialek, M., & Pennycook, G. (2017). The cognitive reflection test is robust to multiple exposures. Behavior Research Methods, 1–7. https://doi.org/10.3758/s13428-017-0963-x

Bluyssen, P. M. (2013). The Healthy Indoor Environment: How to Assess Occupants' Wellbeing in Buildings. *Routledge*. https://doi.org/10.4324/9781315887296

Brañas-Garza, P., Kujal, P., & Lenkei, B. (2019). Cognitive reflection test: Whom, how, when. *Journal of Behavioral and Experimental Economics*, 82. https://doi.org/10.1016/j.socec.2019.101455

Byrne, N. M., Hills, A. P., Hunter, G. R., Weinsier, R. L., & Schutz, Y. (2005). Metabolic equivalent: One size does not fit all. *Journal of Applied Physiology, 99(3)*, 1112–1119. https://doi.org/10.1152/japplphysiol.00023.2004

Byrnes, J. P., Miller, D. C., & Schafer, W. D. (1999). Gender Differences in Risk Taking: A Meta-Analysis. *Psychological Bulletin, 125*(3).

[11] For an overview of CRT and CRT extension recognition and recollection, see appendix table 9

Cao, M., & Wei, J. (2005). Stock market returns: A note on temperature anomaly. *Journal of Banking and Finance, 29*(6), 1559–1573. https://doi.org/10.1016/j.jbankfin.2004.06.028

Carter, E. C., Kofler, L. M., Forster, D. E., & Mccullough, M. E. (2015). A Series of Meta-Analytic Tests of the Depletion Effect: Self-Control Does Not Seem to Rely on a Limited Resource *APA NLM. Association, 144*(3), 0. https://doi.org/10.1037/xge0000083.supp

Chang, T. Y., & Kajackaite, A. (2019). Battle for the thermostat: Gender and the effect of temperature on cognitive performance. *PLoS ONE, 14*(5). https://doi.org/10.1371/journal.pone.0216362

Cheema, A., & Patrick, V. M. (2012). Influence of Warm Versus Cool Temperatures on Consumer Choice: A Resource Depletion Account. *Journal of Marketing Research, 49*(6), 984–995. https://doi.org/10.1509/jmr.08.0205

Craig, C., Overbeek, R. W., Condon, M. V., & Rinaldo, S. B. (2016). A relationship between temperature and aggression in NFL football penalties. *Journal of Sport and Health Science, 5*(2), 205–210. https://doi.org/10.1016/j.jshs.2015.01.001

Cunningham, M. R., & Baumeister, R. F. (2016). How to Make Nothing Out of Something: Analyses of the Impact of Study Sampling and Statistical Interpretation in Misleading Meta-Analytic Conclusions. *Frontiers in Psychology, 7*(1639). https://doi.org/10.3389/fpsyg.2016.01639

Denson, T. F., DeWall, C. N., & Finkel, E. J. (2012). Self-control and aggression. *Current Directions in Psychological Science, 21*(1), 20–25. https://doi.org/10.1177/0963721411429451

Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., & Wagner, G. G. (2011). Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association, 9*(3), 522–550. https://doi.org/10.1111/j.1542-4774.2011.01015.x

Eichholtz, P., Holtermans, R., Kok, N. (2019). Environmental Performance of Commercial Real Estate: New Insights into Energy Efficiency Improvements. *The Journal of Portfolio Management, 45*(7), 113-129.

Engelbert, M., & Carruthers, P. (2010). Introspection. *Wiley Interdisciplinary Reviews: Cognitive Science, 1*(2), 245–253). https://doi.org/10.1002/wcs.4

Engle, R. W., & Kane, M. J. (2003). Executive Attention, Working Memory Capacity, and a Two-Factor Theory of Cognitive Control. *Psychology of Learning and Motivation - Advances in Research and Theory, 44*, 145–199. https://doi.org/10.1016/S0079-7421(03)44005-X

Evans, J. S. (2007). On the resolution of conflict in dual-process theories of reasoning. *Thinking and Reasoning, 13*(4), 321–329.

Falk, A., Dohmen, T., & Huffman, D. (2016). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *IZA Discussion Paper*, No. 9674(9674). http://ftp.iza.org/dp9674.pdf

Frederick, S. (2005). Cognitive Reflection and Decision Making. *Journal of Economic Perspectives, 19*(4), 25–42. https://doi.org/10.1257/089533005775196732

Frey, R., Pedroni, A., Mata, R., Rieskamp, J., & Hertwig, R. (2017). Risk preference shares the psychometric structure of major psychological traits. *Science Advances, 3*(10). https://doi.org/10.1126/sciadv.1701381

Gawronski, B., & Bodenhausen, G. V. (2006). Associative and propositional processes in evaluation: An integrative review of implicit and explicit attitude change. *Psychological Bulletin, 132*(5), 692–731. https://doi.org/10.1037/0033-2909.132.5.692

Grether, W. F. (1973). Human performance at elevated environmental temperatures. *Aerospace Medicine, 44*(7), 747–755. http://www.ncbi.nlm.nih.gov/pubmed/4715089

Hagger, M. S., Chatzisarantis, N. L. D., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., Brand, R., Brandt, M. J., Brewer, G., Bruyneel, S., Calvillo, D. P., Campbell, W. K., Cannon, P. R., Carlucci, M., Carruth, N. P., Cheung, T., Crowell, A., De Ridder, D. T. D., Dewitte, S., . . . Zwienenberg, M. (2016). A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspectives on Psychological Science, 11*(4), 546–573. https://doi.org/10.1177/1745691616652873

Hancock, P. A., & Vasmatzidis, I. (1998). Human occupational and performance limits under stress: The thermal environment as a prototypical example. *Ergonomics, 41*(8), 1169–1191. https://doi.org/10.1080/001401398186469

Hancock, P. A., & Vasmatzidis, I. (2003). Effects of heat stress on cognitive performance: The current state of knowledge. *International Journal of Hyperthermia, 19*(3), 355–372. https://doi.org/10.1080/0265673021000054630

Hermalin, B. E., & Weisbach, M. S. (1991). The Effects of Board Composition and Direct Incentives on Firm Performance. *Financial Management, 20*(4), 101. https://doi.org/10.2307/3665716

Holt, C. A., & Laury, S. K. (2002). Risk Aversion and Incentive Effects. *American Economic Review, 92*(5), 1644–1655. https://doi.org/10.1257/000282802762024700

Huizenga, C., Abbaszadeh, S., Zagreus, L., & Arens, E. (2006). Air quality and thermal comfort in office buildings: Results of a large indoor environmental quality survey. *Proceedings of Healthy Buildings, 3*, 393–397. https://doi.org/10.12659/PJR.894050

Kahneman, D. (1973). *Attention and Effort*. Englewood Cliffs.

Kingma, B., & Van Marken Lichtenbelt, W. (2015). Energy consumption in buildings and female thermal demand. *Nature Climate Change, 5*(12), 1054–1056. https://doi.org/10.1038/nclimate2741

Künn, S., Palacios, J., & Pestel, N. (2019). The Impact of Indoor Climate on Human Cognition: Evidence from Chess Tournaments. *IZA Discussion Paper*, 12632.

Lan, L., Lian, Z., Pan, L., & Ye, Q. (2009). Neurobehavioral approach for evaluation of office workers' productivity: The effects of room temperature. *Building and Environment, 44*(8), 1578–1588. https://doi.org/10.1016/j.buildenv.2008.10.004

Leith, K. P., & Baumeister, R. F. (1996). Why do bad moods increase self-defeating behavior? Emotion, risk taking, and self-regulation. *Journal of Personality and Social Psychology, 71*(6), 1250–1267. https://doi.org/10.1037/0022-3514.71.6.1250

MacLeod, C. M. (1991). Half a century of reseach on the stroop effect: An integrative review. *Psychological Bulletin, 109*(2), 163–203. https://doi.org/10.1037/0033-2909.109.2.163

MacNaughton, P., Satish, U., Laurent, J. G. C., Flanigan, S., Vallarino, J., Coull, B., Spengler, J. D., & Allen, J. G. (2017). The impact of working in a green certified building on cognitive function and health. *Building and Environment, 114*, 178–186. https://doi.org/10.1016/j.buildenv.2016.11.041

McDonald, J. H. (2014). *Handbook of Biological Statistics (3rd ed.).* Sparky House Publishing.

Meyer, A., Zhou, E., & Frederick, S. (2018). The non-effects of repeated exposure to the cognitive reflection test. *Judgment and Decision Making, 13*(3).

Muraven, M., & Baumeister, R. F. (2000). Self-Regulation and Depletion of Limited Resources: Does Self-Control Resemble a Muscle? *Psychological Bulletin, 126*(2), 247–259. https://doi.org/10.1037/0033-2909.126.2.247

Palacios, J., Eichholtz, P., Kok, N. (2020). Moving to productivity: The benefits of healthy buildings. *PLoS ONE, 15*, https://doi.org/10.1371/journal.pone.0236029

Pérez-Lombard, L., Ortiz, J., & Pout, C. (2008). A review on buildings energy consumption information. *Energy and Buildings, 40*(3), 394–398. https://doi.org/10.1016/j.enbuild.2007.03.007

Perry, J. L., & Porter, L. W. (1982). Factors Affecting the Context for Motivation in Public Organizations. *Academy of Management Review, 7*(1), 89–98. https://doi.org/10.5465/amr.1982.4285475

Satish, U., Mendell, M. J., Shekhar, K., Hotchi, T., Sullivan, D., Streufert, S., & Fisk, W. J. (2012). Is CO2 an indoor pollutant? direct effects of low-to-moderate CO2 concentrations on human decision-making performance. *Environmental Health Perspectives, 120*(12), 1671–1677. https://doi.org/10.1289/ehp.1104789

Shibasaki, M., Namba, M., Oshiro, M., Kakigi, R., & Nakata, H. (2017). Suppression of cognitive function in hyperthermia; From the viewpoint of executive and inhibitive cognitive processing. *Scientific Reports*, 7. https://doi.org/10.1038/srep43528

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018). Performance on the cognitive reflection test is stable across time. *Judgment and Decision Making, 13*(3). https://doi.org/10.2139/ssrn.3115809

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and easy implementation of the Benjamini-Hochberg procedure for controlling the false positive rate in multiple comparisons. *Journal of Educational and Behavioral Statistics, 27*(1), 77–83. https://doi.org/10.3102/10769986027001077

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory and Cognition, 39*(7), 1275–1289. https://doi.org/10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking and Reasoning, 20*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Van Lange, P. A. M., Rinderu, M. I., & Bushman, B. J. (2017). Aggression and violence around the world: A model of CLimate, Aggression, and Self-control in Humans (CLASH). *Behavioral and Brain Sciences, 40*. https://doi.org/10.1017/S0140525X16000406

Van Ooijen, A. M. J., Van Marken Lichtenbelt, W. D., Van Steenhoven, A. A., & Westerterp, K. R. (2004). Seasonal changes in metabolic and temperature responses to cold air in humans. *Physiology and Behavior, 82*(2–3), 545–553. https://doi.org/10.1016/j.physbeh.2004.05.001

Vermeulen, W. J. V., & Hovens, J. (2006). Competing explanations for adopting energy innovations for new office buildings. *Energy Policy, 34*(17), 2719–2735. https://doi.org/10.1016/j.enpol.2005.04.009

Wageman, R., & Baker, G. (1997). Incentives and cooperation: The joint effects of task and reward interdependence on group performance. *Journal of Organizational Behavior, 18*(2), 139–158. https://doi.org/10.1002/(SICI)1099-1379(199703)18:2<139::AID-JOB791>3.0.CO;2-R

Wang, X. (2017). An Empirical Study of the Impacts of Ambient Temperature on Risk Taking. *Psychology, 08*(07), 1053–1062. https://doi.org/10.4236/psych.2017.87069

Welsh, M. B., Burns, N. R., & Delfabbro, P. H. (2013). The Cognitive Reflection Test: how much more than Numerical Ability? *35th Annual Conference of the Cognitive Science Society, 35*(35), 1587–1592.

Wright, K. P., Hull, J. T., & Czeisler, C. A. (2002). Relationship between alertness, performance, and body temperature in humans. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology, 283*(6), R1370–R1377. https://doi.org/10.1152/ajpregu.00205.2002.-Body

Wyon, D. P. (1974). The effects of moderate heat stress on typewriting performance. *Ergonomics, 17*(3), 309–317. https://doi.org/10.1080/00140137408931356

Zhang, D. C., Highhouse, S., & Rada, T. B. (2016). Explaining sex differences on the Cognitive Reflection Test. *Personality and Individual Differences, 101*, 425–427. https://doi.org/10.1016/j.paid.2016.06.034

Zhang, F., & De Dear, R. (2017). University students' cognitive performance under temperature cycles induced by direct load control events. *Indoor Air, 27*(1), 78–93. https://doi.org/10.1111/ina.12296

Zhang, F., Haddad, S., Nakisa, B., Rastgoo, M. N., Candido, C., Tjondronegoro, D., & de Dear, R. (2017). The effects of higher temperature setpoints during summer on office workers' cognitive load and thermal comfort. *Building and Environment, 123*, 176–188. https://doi.org/10.1016/j.buildenv.2017.06.048

Zhang, X., Wargocki, P., Lian, Z., & Thyregod, C. (2017). Effects of exposure to carbon dioxide and bioeffluents on perceived air quality, self-assessed acute health symptoms, and cognitive performance. *Indoor Air, 27*(1), 47–64. https://doi.org/10.1111/ina.12284

Zivin, J. G., & Neidell, M. (2012). The impact of pollution on worker productivity. *American Economic Review, 102*(7), 3652–3673. https://doi.org/10.1257/aer.102.7.3652

Appendix

Table 3: Sample Descriptive Statistics

					Male (43%)				Female (57%)				
	Mean (sd)	Min	Max	N	Mean (sd)	Min	Max	N	Mean (sd)	Min	Max	N	*p*-value
Age	21.57 (2.41)	17	31	257	21.70 (2.36)	19	31	119	21.46 (2.45)	17	31	138	*0.34*
Math Proficiency (1-100 scale)	62.97 (17.9)	1	100	257	67.48 (16.42)	2	100	119	59.07 (18.26)	1	88	138	*0.00****
Thermostat Preference (°C, in winter)	21.91 (2.65)	12	28	235	20.94 (2.74)	12	28	106	21.58 (2.55)	12	27	129	*0.02**

Note. Statistics presented are mean values and standard deviation are presented in parentheses. Extreme thermostat preferences were excluded (below zero degrees and above 30 degrees Celsius). p-values results from nonparametric independent sample t-tests. * indicates p-vale < .05, ** a p-value <.01, and *** a p-value <.001

TABLE 4: Indoor Conditions Descriptive Statistics

	Control	Hot	*p*-value
Indoor Temperature Average (°C)	22.42	28.33	*.000****
Indoor Temperature during Task (°C)	22.60	28.65	*.000****
Indoor CO_2 (ppm)	692.12	726.93	*.722*
Indoor humidity (%)	48.87	39.06	*.003***
Outdoor (°C) temperature at start of the experiment	13.88	14.65	*.656*
Average outdoor (°C) temperature (three days average)	14.44	13.84	*.785*

Note. Statistics presented are mean values and standard deviation are presented in parentheses. ppm stands for particles per million. p-values results from parametric independent sample t-tests. * indicates p-vale < .05, ** a p-value <.01, and *** a p-value <.001

Table 5: Additional Domain-Specific Risk Measures

				Men			Women		
	Control	Hot	*p*-value	Control	Hot	*p*-value	Control	Hot	*p*-value
Panel A. Self-reported Risk Attitude									
General	5.77 (1.91)	5.43 (1.75)	.12	6.08 (1.80)	5.40 (1.77)	.03*	5.49 (2.00)	5.46 (1.74)	.97
Driving	3.39 (2.42)	3.16 (2.38)	.35	4.20 (2.50)	3.48 (2.47)	.07	2.68 (2.13)	2.87 (2.29)	.70
Financial Matters	5.31 (2.24)	5.11 (2.18)	.47	5.80 (2.38)	5.73 (2.23)	.88	4.88 (2.03)	4.57 (2.00)	.31
Sports and Leisure	7.85 (2.13)	7.65 (2.38)	.58	7.93 (1.95)	7.52 (2.35)	.37	7.77 (2.28)	7.77 (2.41)	.95
Work	6.72 (2.18)	6.45 (2.16)	.20	7.03 (1.99)	6.18 (2.06)	.02*	6.45 (2.32)	6.68 (2.23)	.67
Health	4.64 (2.75)	4.17 (2.67)	.18	4.73 (2.41)	4.28 (2.72)	.23	4.57 (3.03)	4.07 (2.65)	.49
Others (social)	6.55 (2.53)	6.58 (2.56)	.98	6.43 (2.27)	5.85 (2.52)	.23	6.65 (2.75)	7.22 (2.44)	.30
Observations	*129*	*128*		*60*	*59*		*69*	*69*	

Note: All scores are on 1-10 likert scale, and all scores are recoded such that 1 is risk averse, and 10 is risk loving. Significance levels are based on nonparametric analysis. Standard deviation are given in parentheses. * indicates p-vale < .05, ** a p-value <.01, and *** a p-value <.001

Table 6: Correlation Table between the Risk Attitude Measure and the Risk Behaviour Measure

	Full sample			Control			Hot		
	M	*SD*	*1*	*M*	*SD*	*1*	*M*	*SD*	*1*
1. Risk Elicitation Task (Holt & Laury, 2002)	6.05	1.91		6.11	1.76		6	2.05	
2. General Risk Attitude (Dohmen et al., 2011)	5.65	1.83	-.12 [-.25, .01]	5.78	1.90	-.05 [-.24,.13]	5.52	1.77	-.20* [-.37,-.01]
Observations	*224*			*111*			*113*		

Note. The Risk Elicitation task has missing values, the summary statistics excluded all risk attitude cases that are matched to missing values for the risk task. Correlation coefficient presented is the Spearman's rho and 95% confidence interval in brackets. * indicates p-vale < .05, ** a p-value <.01, and *** a p-value <.001

Table 7: Multiple Testing Correction Panel A and Panel C for 15% False Discovery Rate level

			Men		Women	
	p-value	Q = 15%	*p*-value	Q = 15%	*p*-value	Q = 15%
Panel A. Self-reported Indoor Variables Satisfaction and Hinder						
Temperature Satisfaction	.00	Sig	.00	Sig	.16	
Air Quality Satisfaction	.00	Sig	.00	Sig	.00	Sig
Light Satisfaction	.07	Sig	.56		.02	Sig
Noise Satisfaction	.18		.58		.18	
Clothing Satisfaction	.14		.02	Sig	.81	
Temperature Hinder	.00	Sig	.00	Sig	.07	Sig
Air Quality Hinder	.00	Sig	.00	Sig	.00	Sig
Light Hinder	.77		.37		.20	
Noise Hinder	.04	Sig	.52		.03	Sig
Clothing Hinder	.89		.17		.30	
Panel C. Self-reported Risk Attitude						
Driving	.35		.07		.70	
Financial Matters	.47		.88		.31	
Sports and Leisure	.58		.37		.95	
Work	.20		.02	Sig	.67	
Health	.18		.23		.49	
Others (social)	.98		.23		.30	

Note. The p-value are the result of nonparametric ranksum tests as shown in table 3. The chosen levels of False Discovery Rates (Q) are chosen given that Q=15% implies less than 1 FDR per 7 tests. Q=5% is the most conservative FDR rate, with the highest risk of False Negatives (McDonald, 2014). Applying the FDR formula (False Discovery Rate = Expected (False Positive / (False Positive + True Positive))) to the risk domain entails that the change of two significant findings amongst 7 domains would be 28.6%. We find two significant findings (in the male sample) if we correct for a FDR as low as 15%. The significance of the general risk attitude in the male sample is robust against a FDR of 12%.

Table 8: Critical Value for 15% False Discovery Rate (Q) per Rank Used for Multiple Testing Correction

False Discovery Rate of 15 %		
Rank	7 items	10 items
1	0,025	0,015
2	0,050	0,030
3	0,075	0,045
4	0,100	0,060
5	0,125	0,075
6	0,150	0,090
7		0,105
8		0,120
9		0,135
10		0,150

Note. The critical p-value thresholds according to the Benjamini & Hochberg (1995) are dependent on the total amount of multiple tests. According to their rank, each level of significance will be compared to their rank critical value as stated in this table. The 7 items critical value are applied to the Self-Reported Risk Attitude (table 5, panel C), the 10 items critical values are applied to the Self-reported Indoor Variables Satisfaction and Hinder (table 5, panel A).

Table 9: Overview of Recognition and Answer Remembering for the CRT Classic and CRT Extention

		Recognize Question	Remember the Answer *		
		%	%		
		Yes	Yes	No	Unsure
CRT Original	Lily pads	45,52	42,54	44,03	13,43
	Widget problem	26,85	26,04	59,38	14,58
	Bat and ball	40,86	45,30	38,46	16,24
CRT Extended	Class ranking	2,72	6,38	78,72	14,89
	Stock market	5,45	58,33	16,67	25,00
	Barrel of water	10,89	7,02	77,19	15,79
	Observations	257			

Note. *The percentage in the remembering column is conditional on recognition. For example: For the Lily pads, of the 45.52% that recognizes the questions, 44.03 % does not remember the answer.

Table 10: Post-Hoc Sensitivity Analysis

		Sample Size		Effect Size *d*	
		Control	Hot	Non-Parametric Mann-Whitney	Parametric T-Test
Majority Measures	Full Sample	129	128	0,42	0,41
	Men	60	59	0,62	0,61
	Women	69	69	0,58	0,56
Risk Elicitation Task	Full Sample	111	113	0,45	0,44
	Men	53	51	0,66	0,65
	Women	58	62	0,62	0,60

Note. Effect size sensitivity is reported per groupsize. The first rows apply to the majority of all presented results in the paper. Only for the risk elicitation task, the latter rows apply, due to some exclusion criteria applied in that sample. We present for each sample-size sensitivity estimates for both parametric as well as non-parametric tests.