Potential Sex Difference in the Effects of Mild Acute Stress on Executive Functions

by

Haolu Zhang

B.Sc., The University of British Columbia, 2013

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Neuroscience)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

October 2016

© Haolu Zhang, 2016
Abstract

Prefrontal cortex (PFC)-dependent executive functions (EFs) are critical for reasoning, problem-solving, self-control and planning. The PFC dopamine (DA) level has been demonstrated to modulate EFs in an inverted U-shaped curve, where an intermediate level of DA is optimal. Unlike in other brain regions, PFC DA systems: 1) relies highly on the catechol-O-methyltransferase (COMT) enzyme for clearing released DA; and 2) can be activated even by mild stress. Estradiol (E2) has been shown to down-regulate COMT gene transcription, causing the activity of COMT enzyme to be ~30% less in women than in men.

Animal studies have repeatedly shown that stress facilitates cognitive functions dependent on the hippocampus and/or PFC in males, but impairs them in females. Therefore, based on Diamond’s hypothesis that baseline PFC DA levels are higher and closer to the optimal level in women during menstrual phases when their circulating E2 are elevated, than in men, we predicted that mild stress would facilitate EF performance in men but impair it in women when their circulating E2 levels are high.

In a crossover design, healthy young adults (both men and women), all COMT Val\textsuperscript{158}Met heterozygotes, were each tested twice (once with social-evaluative stress and once without, order counterbalanced) on five EF tasks which tapped on the core EFs of inhibitory control, working memory and cognitive flexibility, and one higher-level EF, reasoning. Women were randomly assigned to the low-E2 (F-L) group or high-E2 (F-H) group. Women
in the F-L group were tested during the early follicular phase (low E2 level). Women in the F-H group were tested during the midluteal phase (high E2 level).

Our social-evaluative procedure was showed to succeed in inducing physiological and subjective stress responses and significantly impaired the performance of the F-H group on one index of inhibitory control, whereas the performance of the M and F-L groups showed a trend towards enhancement. Similar trends (M and F-L: stress-induced enhancement; F-H: stress-induced impairments) were found for some other indices in the first two tasks. These results emphasized that the ways of improving EFs need to be considered in a sex-specific and hormone-dependent manner.
Preface

The research in this thesis was designed and initiated by Prof. Adele Diamond with the assistance of Prof. Elizabeth Hampson, Prof. Clemens Kirschbaum, Prof. Weihong Song and graduate student, Golnoush Alamian. The majority of the research, including subject recruitment, intake interviews, subject scheduling, subject testing, questionnaires and interviews, preparation of saliva samples for the various assays, data entry, cleaning data, data analyses, subject retention for subsequent sessions, preparation of posters and training research assistants, was conducted by the author of this thesis, with suggestions from Prof. Adele Diamond. Prof. Hampson’s lab conducted the hormonal assays. Prof. Kirschbaum’s lab performed the cortisol and DHEA immunoassays. Prof. Weihong Song’s lab performed the COMT genotyping. David Abbott helped a lot on the methods part. Prof. Rich White helped a lot on the statistical analyses part. Research Assistants in the Developmental Cognitive Neuroscience lab, including Kristina Balce, Arminder Chandi, and Shahab Zareyan, helped greatly on subject recruitment, subject testing and data entry. Prof. Adele Diamond helped greatly on revisions of this thesis.

All experimental procedures in this work were reviewed and approved by the UBC Clinical Research Ethics Board (project title: Differences by Sex and Genotype in the Effects of Stress on Executive Functions; certificate #H011-00739).
# Table of Contents

Abstract............................................................................................................................ ii

Preface.................................................................................................................................. iv

Table of Contents ....................................................................................................................... v

List of Tables .......................................................................................................................... ix

List of Figures ........................................................................................................................ x

Acknowledgments ................................................................................................................... xiii

Glossary .................................................................................................................................. xiv

1. Introduction ......................................................................................................................... 1

2. Methods and Materials ...................................................................................................... 9

   2.1 Participants ..................................................................................................................... 9

   2.2 Procedure ..................................................................................................................... 11

   2.3 More stressful and calmer conditions ....................................................................... 13

   2.4 Endocrine and autonomic assessments ................................................................ 14

   2.5 COMT genotyping .................................................................................................... 17
2.6 Subjective assessment of stress over the month preceding testing: The Perceived Life Stress Scale ................................................................. 17

2.7 Subjective assessment of stress immediately before and after testing .... 18

2.8 Cognitive tests .................................................................................. 19

2.8.1 Flanker/Reverse Flanker task: .................................................... 20

2.8.2 Hearts and Flowers (or Dots) task: ............................................. 25

2.8.3 Raven’s Advanced Progressive Matrices (RAPM): ....................... 31

2.8.4 Re-Ordering Digit Span task: ..................................................... 32

2.8.5 Visuo-Spatial Working Memory Span task: ................................. 35

2.9 Statistical analyses ........................................................................ 37

3. Results ............................................................................................... 41

3.1 Demographic characteristics ............................................................ 41

3.2 Sex hormone assessments ............................................................... 41

3.3 Stress levels of participants ............................................................. 44

3.3.1 Blood pressure and heart rate: .................................................. 44
3.3.2 Cortisol and cortisol/DHEA ratio: ................................................................. 52

3.3.3 Subjective assessments of stress: ................................................................. 58

3.4 Executive function (EF) cognitive tests .......................................................... 62

3.4.1 Order effect: .............................................................................................. 62

3.4.2 Flanker/Reverse Flanker task: ................................................................. 62

3.4.3 Heart and Flower (or Dots) task: ............................................................. 70

3.4.4 Raven’s Advanced Progressive Matrices: ................................................. 75

3.4.5 Re-Ordering Digit Span task (verbal WM): ............................................ 77

3.4.6 Visuo-Spatial WM task: ................................................................................ 77

4. Discussion ....................................................................................................... 81

4.1 The effectiveness of our stress induction procedure................................. 81

4.2 Sex differences in stress responsivity ......................................................... 83

4.3 Sex differences in the effects of stress on executive functions..................... 83

4.4 Potential underlying mechanisms................................................................. 89

4.5 Limitations .................................................................................................... 91
4.6 Implications .................................................................................................................................................. 94

5. Conclusion ....................................................................................................................................................... 97

Bibliography .......................................................................................................................................................... 98

Appendix ............................................................................................................................................................... 124
List of Tables

Table 1 Characteristics of participants in the three subject groups. .......................... 41
List of Figures

Figure 1 Illustration of how hypothesized baseline difference in PFC DA levels between males and females would explain the observed sex differences in the effects of mild stress on executive functions in rodents.................................................................8

Figure 2 Timeline for each testing session.................................................................12

Figure 3 Examples of stimuli used in our Flanker/Reverse Flanker task......................21

Figure 4 Examples of stimuli used in our Hearts and Flowers (or Dots) task...............27

Figure 5 A sample item from Raven’s Advanced Progressive Matrices......................32

Figure 6 The computer interface that was used by experimenter for recording in the Re-Ordering Digit Span task........................................................................................................34

Figure 7 An example of grid (A) and the masking stimulus (B) in the Visuo-Spatial Working Memory Span task.................................................................37

Figure 8 Mean estradiol (A) and progesterone (B) levels...........................................43

Figure 9 Results for autonomic indicators of stress.................................................48

Figure 10 Differences in autonomic indicators of stress between the two conditions (more stressful and calmer) for each subject group........................................50

Figure 11 Results for HPA-axis indicators of stress.................................................56
Figure 12 Differences of HPA-axis indicators of stress between two conditions (more stressful and calmer). ................................................................. 58

Figure 13 Mean self-reported stress levels throughout the more stressful and calmer conditions for each sex group. ........................................................................................................... 60

Figure 14 Differences in self-reported stress levels comparing the more stressful and calmer conditions for each subject group. .............................................................. 61

Figure 15 An index assessing inhibitory control (difference in percentage of correct responses on incongruent trials versus easy trials (congruent and no-distractor trials) in Block 2 and 3 combined was divided by percentage of correct response on easy trials) on the Flanker/Reverse Flanker task for each subject group for each condition (more stressful and calmer; see Panel A) and differences scores for this index (i.e. index in the calm condition minus index in the stress condition) for the F-H group and the two other subject groups combined (Panel B). ......................... 66

Figure 16 Mean difference scores (index in the calm condition minus index in the stress condition) on the Flanker/Reverse Flanker task for Global Switch Cost-1 (calculated from mean RTs; Panel A) and difference scores for Global Switch Cost-2 (calculated from mean RTs; Panel B) and composite scores for inhibitory control (calculated from mean RTs; Panel C) and composite scores for inhibitory control and cognitive flexibility (calculated from mean RTs; Panel D). ................. 70
Figure 17 Mean difference scores for RT difference between incongruent trials and congruent trials in Block 3 (Panel A) and RT difference between incongruent trials in a switching environment (Block 3) and incongruent trials in an environment where the rule did not change (Block 2; Panel B), Local Switch Cost for RT in Block; Panel C), and Global Switch Cost-1 for RT (Panel D) and composite scores for inhibitory control for the Heart and Flower (or Dots) task (calculated from mean RT; panel E) and composite scores for cognitive flexibility for the Heart and Flower (or Dots) task (calculated from mean RT; Panel F) and composite scores for inhibitory control across first two tasks (calculated from mean RT; panel G) and composite scores for cognitive flexibility across first two tasks (calculated from mean RT; Panel H)........................................................................................................................................ 74

Figure 18 Performance (z scores for total scores) on Raven’s Advanced Progressive Matrices by subject group and condition. ........................................................................................................... 76

Figure 19 Mean longest span for the Re-Ordering Digit Span task (A) and our Visuo-Spatial WM task (B) and composite scores across these two tasks (C). ................. 80
Acknowledgments

First and foremost, I wish to express my great appreciation to my supervisor, Prof. Adele Diamond, for her persistent support, encouragement and deep insight in the research area throughout my Master’s study. I would also like to thank Prof. Elizabeth Hampson for providing hormonal analyses and for her useful guidance and suggestions on interpreting hormonal data. I would like to thank my thesis examination committee members for their time and suggestions.

Many thanks go to David Abbott, who provided me a lot of help on many aspects of working in the Developmental Cognitive Neuroscience lab. I would like to thank all my labmates, Daphne Ling, Alyssa Ash, Kristina Balce, Shahab Zareyan, and Arminder Chandi. Thanks a lot for their help and useful feedbacks. I also would like to thank my friends, Dr. Aiping Liu, Yue Wang, Jessica Guo, Xiaowei Tao, Anna Luo, Ruixuan Shi, Dr. Xun Chen, Shumin Wang, and many others at UBC for the friendship and moral support they provided.

I owe my deepest gratitude to my parents for their endless unconditional love and immense support from thousands of miles away. A special thank goes to my husband, Yu Su, for his love, understanding, and for sharing the joys, happiness as well as frustrations with me.
<table>
<thead>
<tr>
<th>Term</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANOVA</td>
<td>Analysis of variance</td>
</tr>
<tr>
<td>COMT</td>
<td>catechol-O-methyltransferase</td>
</tr>
<tr>
<td>DA</td>
<td>Dopamine</td>
</tr>
<tr>
<td>E2</td>
<td>Estradiol</td>
</tr>
<tr>
<td>EF</td>
<td>Executive functions</td>
</tr>
<tr>
<td>GC</td>
<td>Glucocorticoid</td>
</tr>
<tr>
<td>GR</td>
<td>Glucocorticoid receptors</td>
</tr>
<tr>
<td>GLMM</td>
<td>Generalized linear mixed effects model</td>
</tr>
<tr>
<td>HPA</td>
<td>Hypothalamic-pituitary-adrenal</td>
</tr>
<tr>
<td>LME</td>
<td>Linear mixed effects model</td>
</tr>
<tr>
<td>Met</td>
<td>Methionine</td>
</tr>
<tr>
<td>NE</td>
<td>Norepinephrine</td>
</tr>
<tr>
<td>OVX</td>
<td>Ovariectomized</td>
</tr>
<tr>
<td>P4</td>
<td>Progesterone</td>
</tr>
<tr>
<td>PET</td>
<td>Positron emission tomography</td>
</tr>
<tr>
<td>PFC</td>
<td>Prefrontal cortex</td>
</tr>
<tr>
<td>PSS</td>
<td>Perceived Life Stress Scale</td>
</tr>
<tr>
<td>RAPM</td>
<td>Raven’s Advanced Progressive Matrices</td>
</tr>
<tr>
<td>RT</td>
<td>Reaction Time</td>
</tr>
<tr>
<td>SNS</td>
<td>Sympathetic nervous system</td>
</tr>
<tr>
<td>TSST</td>
<td>Trier Social Stress Test</td>
</tr>
<tr>
<td>Val</td>
<td>Valine</td>
</tr>
<tr>
<td>WM</td>
<td>Working memory</td>
</tr>
</tbody>
</table>
1. Introduction

Executive functions (EFs; also known as “executive control” or “cognitive control”) represent a group of top-down mental functions needed for reasoning, problem-solving, and planning, which are critical for success in school and in life (Diamond, 2013; Jacques & Marcovitch, 2010). It is generally agreed that the complex or higher-order EFs, like logical reasoning, are built from three interrelated core EFs, inhibitory control, working memory (WM), and cognitive flexibility (Diamond, 2013; Logue & Gould, 2013; Miyake et al., 2000). Inhibitory control, including self-control and interference control (attentional control and cognitive inhibition), enables control over one’s behavior, attention, thoughts, and emotions to resist temptations and not act impulsively, and to voluntarily ignore irrelevant information to shield current goals. The second core EF, WM, involves temporary storage and manipulation of information. Third, cognitive flexibility, is being able to switch between different ways of categorizing or perceiving the same thing or solving problems in new ways (see Diamond 2013 for a detailed review).

Prefrontal cortex (PFC), especially lateral PFC, and other brain regions with which it is interconnected subserve EFs (Braver et al., 2002; Zanto et al., 2011). Catecholamines, both dopamine (DA) and norepinephrine (NE), in PFC play a critical role in EFs and have been intensively studied since the landmark discovery by Brozoski et al. (1979), showing the essential role of catechoamines for optimal PFC function. PFC DA modulates EFs in an
inverted U-shaped curve, where an intermediate level of DA is optimal (Cerqueira et al., 2007a; Robbins & Arnsten, 2009; Roth et al., 1988; Vijayraghavan et al., 2007). To achieve optimal EFs, factors that affect PFC DA levels should be considered.

Unlike other brain regions (e.g., the striatum), PFC has limited DA transporter (DAT) (Durston et al., 2005; Lewis et al., 2001; Sesack et al., 1998), which provides the best mechanism for clearing released DA. PFC, therefore, relies heavily on the catechol-O-methyltransferase (COMT) enzyme to metabolize released DA (Matsumoto et al., 2003), whereas regions with sufficient DAT are far less dependent on the COMT enzyme. The COMT enzyme catalyzes the transfer of a methyl group to DA to inactivate it. One common polymorphism of the COMT gene is a single nucleotide substitution of methionine (Met) for Valine (Val) at codon 158 (Val^{158}Met) that leads to a ~35-50% decrease in the activity of the enzyme, thereby allowing more DA to remain longer in PFC (Chen et al., 2004; Lachman et al., 1996). A growing body of evidence suggests that individuals homozygous for COMT-Met^{158} have better inhibitory control, WM and greater PFC efficiency than individuals homozygous for COMT-Val^{158}, while individuals with the Val/Met genotype are intermediate (e.g. (Barnett et al., 2007; Blasi et al., 2005; Bruder et al., 2005; Diamond et al., 2004; Egan et al., 2001; Malhotra et al., 2002). Individuals homozygous for COMT-Met^{158} show enhanced performance on tasks requiring inhibitory control and WM, also known as “cognitive stability” (Bilder et al., 2004; Blasi et al., 2005; Bruder et al., 2005; Egan et al., 2001; Nolan et al., 2004). Conversely, individuals homozygous for COMT-Val^{158} have
sometimes been found to perform better on tasks requiring cognitive flexibility (Colzato et al., 2010; Nolan et al., 2004). However, the relation between COMT polymorphism and EFs is not completely clear, with some inconsistent results across studies (Blanchard et al., 2011; Dennis et al., 2010).

The relation between COMT polymorphisms and EFs may be modulated by sex hormones. In female rodents, DA release increases with natural elevations in estrogens during the estrous cycle and after exogenous administration of estrogens (Dazzi et al., 2007; Xiao & Becker, 1994). In humans, several studies report that estradiol (E2; one of major estrogens) down-regulates COMT gene transcription (Ho et al., 2008; Jiang et al., 2003) which causes the activity of the COMT enzyme in PFC to be ~30% lower in women than in men (Chen et al., 2004). Some studies have investigated the role of E2 on EFs, such as WM, over the past two decades but the data have been inconsistent (e.g. (Evans & Hampson, 2015; Hampson, 1995; Lac reuse, 2006; Pompili et al., 2012)). Significant effects of E2 or the menstrual cycle on EFs have been reported in some studies (Colzato et al., 2012; Colzato et al., 2010; Gasbarri et al., 2008; Hampson & Morley, 2013a) but not in others (Bannbers et al., 2012; Solis-Ortiz & Corsi-Cabrera, 2008). There is evidence from both animal and human studies showing that compared to females when their E2 levels are low and males, females when their E2 levels are high show better performance on PFC-dependent tasks, including EFs (Hampson & Morley, 2013b; Shors & Leuner, 2003). Diamond (2007) was the first to suggest the roles of the baseline DA level and COMT genotype in studying sex differences in
or estrogenic effects on PFC-dependent cognitive functions. Jeanette Evans in her lab demonstrated that when E2 levels are high (midluteal phase) women homozygous for \textit{COMT}-Val\textsuperscript{158} show better EFs than women homozygous for \textit{COMT}-Met\textsuperscript{158} (Evans et al., 2009). Jacobs and D'Esposito (2011) investigated the direction of estrogenic effects on WM using \textit{COMT} genotype as the index of baseline PFC DA levels in healthy young women, and showed that when E2 levels were high (in late follicular phase), women homozygous for \textit{COMT}-Val\textsuperscript{158} performed better on the n-back task, whereas when E2 levels were low (in early follicular phase), the better performers were women homozygous for \textit{COMT}-Met\textsuperscript{158}.

Accumulating evidence shows that even mild stress increases DA release in PFC, but not in the striatum or elsewhere in the brain, and this increase of DA in PFC has a critical impact on EFs (Arnsten, 2009b; Cerqueira et al., 2007b; Kodama et al., 2014; Lataster et al., 2011; Nagano-Saito et al., 2013; Roth et al., 1988). These stress effects on PFC DA levels translate well from rodent to monkey to human (Arnsten, Wang, et al., 2015). Several recent studies have consistently shown that acute, short-term psychological stress increases DA release in PFC of both men and women by using DA D2/3 positron emission tomography (PET) in independent samples of healthy individuals (Vaessen et al., 2015).

Exposure to stress activates stress circuits in the nervous system, which leads to immediate activation of the sympathetic nervous system (SNS), hypothalamic-pituitary-adrenal (HPA) axis and increased firing of central catecholaminergic neurons (Arnsten,
The release of many neurochemical messengers, especially catecholamine and glucocorticoids (GCs; cortisol in humans), are then increased. GCs can cross the blood-brain barrier and reach all brain regions. PFC contains the highest density of GC receptors in the brains of primates (Sanchez et al., 2000). In male rodents, glucocorticoid receptors (GRs) within the mPFC have also been demonstrated to regulate the stress-induced increase of PFC DA levels (Butts et al., 2011).

For almost 20 years, studies in animals have shown that stress differentially affects cognitive performance on PFC-dependent tasks in males and females (Andreano & Cahill, 2006; Arnsten & Goldman-Rakic, 1998; Bangasser & Shors, 2004; Shors et al., 1998; Wood et al., 2001). For example, Shors’ lab has demonstrated that stress facilitates trace eyeblink conditioning (which depends on both the hippocampus and PFC (Gilmartin & McEchron, 2005)) in male rats, but impairs it in female rats (Shors et al., 1998; Wood et al., 2001). Delayed eyeblink conditioning, which does not depend on the hippocampus or PFC, does not show this sex difference in the effect of stress. Most of these studies focused on the modulation by GCs of the hippocampus to explain the stress-related sex differences in cognition. About 12 years ago, research coming out of Arnsten’s lab started to focus on the role of estrogens and PFC in mediating this sex difference. Their rodent studies showed that females in the proestrus phase, when levels of estrogens are highest, are impaired by moderate restraint stress or by a low dose of an anxiogenic drug (a benzodiazepine inverse
agonist [FG7142] that simulates the feeling of being stressed), whereas males, and females in
the estrus phase (when estrogen levels are low) show no effect or a trend toward
improvement after the stressor (Shansky et al., 2004; Shansky et al., 2006). The role of E2 in
mediating impairing effects of stress on WM performance in female has been further
supported by studies in ovariectomized (OVX) rats. OVX female rats with estradiol
replacements show WM impairments that are similar to those of females in proestrus, while
OVX female rats with placebo perform like males (Shansky et al., 2004).

In humans, most studies on stress effects on EFs have not considered sex differences
and have shown inconsistent results (e.g. (Duncko et al., 2009; Oei et al., 2006; Plessow et
al., 2011; Qin et al., 2012; Schoofs et al., 2008; Schoofs et al., 2009). Recently, both
Cornelisse et al. (2011) and Schoofs et al. (2013a) reported the effects of acute stress on WM
performance differed in men and women. After being exposed to the Trier Social Stress Test
(TSST), a well-established social-evaluative stress paradigm, performance (reaction time, RT)
on the n-back task was enhanced in men, but was impaired or did not change in women. It is
worth noting that cortisol levels in men increased significantly, whereas women showed less
or a blunted cortisol response. However, the investigators focused on the effects of cortisol in
isolation without considering genetic influences (e.g., COMT genotype) or baseline DA
levels. As well, their designs were between-subject, not with each participant as his or her
own control. Moreover, E2 levels were not measured in women in their studies, and
menstrual phases were not well controlled.
In the current study, we aimed to test Diamond’s hypothesis that baseline levels of DA in PFC are higher in females at least during the portion of the menstrual cycle when E2 levels are high (due to E2 reducing COMT enzymatic activity) and are closer to optimal in females than in males. We investigated the potential sex difference in the effects of mild stress on EFs in healthy young men and women (see Figure 1). For this purpose, social evaluative stress was chosen as the mild stressor (Dickerson & Kemeny, 2004; Dickerson et al., 2008). Given that 1) E2 seems to increase DA levels in PFC so that women (at least when their E2 levels are high) should have higher baseline DA levels in PFC than men (Chen et al., 2004), 2) mild stress seems to increase DA levels in PFC (give reference) and 3) mild stress facilitates EF performance in males but impairs that in females during the phase of their estrous cycle when E2 levels are high (Shansky et al., 2004; Shansky et al., 2006), we predicted that mild stress would improve men’s EFs but impair women’s EFs when their E2 levels were high.
Figure 1 Illustration of how hypothesized baseline difference in PFC DA levels between males and females would explain the observed sex differences in the effects of mild stress on executive functions in rodents

Adapted from “Biological and social influences on cognitive control processes dependent on prefrontal cortex.” by A. Diamond, 2011, Progress in Brain Research, 189, p. 328. Copyright 2011 by Elsevier B.V. Adapted with permission. *at least when E2 levels are high.
2. Methods and Materials

2.1 Participants

Healthy young adults (N=83; 56 women, 28 tested when their E2 levels were high; 28 tested when their E2 levels were low, and 27 men), all COMT Val-Met heterozygotes aged 21-35 (when sex hormones are at maximal levels and EFs are sharpest) were each tested twice, one month apart (once with social evaluative stress and once without). They were recruited from the University of British Columbia and the greater Vancouver area via posters and advertisements. Monetary compensation was provided for their participation ($10 for the information session, $15 for Session 1 and $25 for Session 2). The eligibility of potential participants was assessed over the phone or through emails.

The inclusion criteria were: 1) No reported EF-testing contra-indications, including serious health problems likely to affect cognition, head trauma or concussion where they lost consciousness, learning disability or mental health disorder, or on any medication that might affect cognition,

2) Not a smoker (due to the effects of nicotine on EFs and the HPA axis (Ernst et al., 2001; Kirschbaum & Hellhammer, 1994)), and

3) No history of major life traumas that would affect EFs or stress responsivity (Aupperle et al., 2012).
4) COMT heterozygotes.

Additional inclusion criteria for women were that they have a relatively regular menstrual cycle, were not pregnant or lactating, and had not taken any hormone-releasing contraceptives for at least four months.

Women were randomly assigned to the low-E2 (F-L) group or high-E2 (F-H) group. Women in the F-L group were tested for both their sessions during the early follicular phase (days 3-5 after the onset of menstruation. Women in the F-H group were tested for both their sessions during the midluteal phase. The midluteal phase was estimated by counting back 6-8 days from the anticipated onset date of the next menstruation. Menstrual phases were validated post hoc by E2 and progesterone (P4) levels, which resulted in 7 female participants (3 in the F-L group and 4 in the F-H group) being excluded because their hormones did not match their supposed menstrual phase (explained in Section 2.4). Additionally, another 6 participants (1 male and 5 in the F-L group) were excluded based on scores of the Perceived Stress Scale indicating that they arrived at one session far more stressed than at their other session (explained in Section 2.6). Therefore, 26 men and 44 women (20 in the F-L group and 24 in the F-H group), a total of 70, from the initial 83 participants, were included in further analyses. All participants gave written informed consent. All procedures were approved by the ethical review boards of UBC and Vancouver Coastal Health.
2.2 Procedure

In this crossover design, all participants received two testing sessions, one in each experimental condition (more stressful and calmer). They were randomly assigned to an order of experimental conditions, and this order was counterbalanced across participants within each group. The interval between two sessions was roughly one month or one menstrual cycle. All sessions were scheduled in mid-afternoon because in general, adults have relatively stable and low cortisol levels then (Weitzman et al., 1971) and have peak cognitive performance then (Hasher et al., 1999). To optimize the purity of the saliva, participants were instructed to refrain from eating, drinking (except plain water), or brushing their teeth for at least one hour prior to testing. No gum or other saliva stimulants were used. All participants first came to the Developmental Cognitive Neuroscience Lab of Prof. Adele Diamond in the Detwiller Pavilion on the UBC campus for an information session prior to Session 1 to learn more about the study, sign a consent form, complete a demographic questionnaire and give a saliva sample for COMT genotyping. Female participants were asked about the length of their menstrual cycle and dates of their last and next (estimated) periods.

On each day of testing, participants again came to the Diamond Lab (one male participant was tested in the UBC Robson Square) and rested in our comfortable reception room for a half hour to give them time to relax and calm down from any stress they might have experienced in getting to the lab or during the day. During this time they also filled out
the Perceived Life Stress Scale (PSS-10 version) questionnaire about stresses during the month immediately preceding testing (Cohen et al., 1983). Thereafter, participants entered the testing room and rated their current stress level, provided saliva samples (for baseline cortisol and DHEA levels and for hormonal levels) and the first measures of their blood pressure (BP) and heart-rate (HR) were taken. Then, they completed the session (five EF tests, multiple saliva samples between EF tasks and after the last, multiple BP and HR readings between EF tasks and after the last, and a post-test stress questionnaire) which took about 1½ hours (see Figure 2).

**Figure 2 Timeline for each testing session.**
2.3 More stressful and calmer conditions

In the stress condition, mild psychosocial stress was induced by a novel social-evaluative threat paradigm we have not seen used before. Participants were informed that two testers would be in the room during their cognitive testing to evaluate their performance. During testing, one male and one female tester with clipboard in hand stood on either side of the participant, appearing to be sternly and concernedly evaluating his/her performance and pretending to be taking notes.

Psychosocial stress in the form of social evaluation has been demonstrated to consistently elicit HPA and autonomic nervous system (ANS) responses in humans (Kajantie & Phillips, 2006; Kudielka & Kirschbaum, 2005). However, there are also marked sex differences in these responses. Adult women, especially in the follicular phase, tend to show lower HPA and autonomic responses than men of the same age, whereas women in the luteal phase tend to show levels of stress response (specifically, free cortisol levels) approaching those of men (Kajantie & Phillips, 2006). In almost all cases, the TSST paradigm has been used as the social evaluative stressor. The TSST requires the participant, with little warning, to give a short speech that the participant composes and perform a mental arithmetic task in front of an audience that usually either looks quite stern, shows little emotion, or looks bored (Kirschbaum et al., 1993). We did not use the TSST paradigm for the following reasons: First, using the TSST would have made our sessions even longer, and of unequal length for the more stressful and calmer conditions. Second, our stress induction procedure allowed the
stress to occur in a very natural, believable way during testing. Third, we were interested in the effects of DA, and the dopaminergic response is triggered immediately after stress onset (Hermans et al., 2014). Participants in our study started EF tasks at the same time that the stressor started, instead of 30-40 mins after the start of the stress induction procedure when effects of cortisol start to develop. Compared to the TSST, our stress induction procedure was milder, which might be important in interpreting our results. In the calm condition, no one was in the testing room looking over participants’ shoulders as they completed the EF tests. Participants were de-briefed at the end of stress session, even if that was their first testing session, because we did not want them coming to their second testing all stressed expecting to be harshly evaluated again (or to be less likely to return for a second testing)\(^1\).

### 2.4 Endocrine and autonomic assessments

Following stress and activation of the HPA axis, cortisol is released into the bloodstream. Only a small fraction of the cortisol released remains unbound or “free” and exerts effects on the brain (Kirschbaum & Hellhammer, 2000). Cortisol levels measured in saliva agree closely with the amount of “free” cortisol in blood and hence with the amount of

\(^1\) The de-briefing used to be delivered at the end of the 2\(^{nd}\) testing session. For subjects with stress order 1(receiving stress session during their first session), 3/4 of them received the de-briefing at the end of the 2\(^{nd}\) testing session. The rest received the de-briefing at the end of stress session, even if that was their first session.
cortisol reaching the brain (Gozansky et al., 2005). There is some evidence that measuring cortisol alone is less accurate than measuring the cortisol to dehydroepiandrosterone (DHEA) ratio (Bauer, 2008; Gallagher & Young, 2002; van Broekhoven & Verkes, 2003). DHEA opposes the action of glucocorticoids and lowers cortisol levels (Kalimi et al., 1994; Kroboth et al., 2003). To assess changes in levels of free cortisol and HPA activity, five saliva samples were collected by participants passively drooling/spitting into Salicap devices (Affinity, Ontario, Canada) to collect unstimulated whole saliva. Those five samples were collected right after the relaxation period (baseline, 10 min before the start of testing) and following EF tasks, roughly 20, 45, 60, and 85 mins after the start of testing. Samples were frozen at -20°C and shipped in dry ice to the Lab of Dr. Clemens Kirschbaum at Technische Universität, Dresden, Germany, one of the foremost experts on the effects of cortisol, where samples were thawed, centrifuged, and assayed using high-sensitivity enzyme immunoassays (IBL).

Stress also triggers activation of the ANS, increasing SNS activity and reducing vagal control of BP and HR. Cardiovascular stress responses thus include increases in BP and HR (Hjortskov et al., 2004). There are individual differences in stress responses. Some individuals show heightened HPA activity and heightened ANS activation, whereas others show only one or the other (Kajantie & Phillips, 2006). Thus, the ANS activity of participants in this study was also assessed. That was done by measuring their HR and BP.
HR and BP readings on a digital blood pressure monitor (Omron HEM-711ACN) were recorded right after the relaxation period (baseline) and following each of the five EF tasks.

To verify levels of E2, P4 and testosterone\(^2\) at the time of testing, saliva samples were collected immediately before the onset of testing, stored at -20°C, and shipped to the Lab of Prof. Elizabeth Hampson at the University of Western Ontario, London, ON, the foremost expert in assaying sex hormone levels from saliva and one of the foremost researchers on sex differences. Saliva assays were used because they offer practical advantages over serum and provide a precise estimate of the bioavailable fraction of the hormones (Becker et al., 2005; Hampson & Young, 2008). Salivary E2 and P4 levels were determined using highly sensitive radioimmunoassay methods that have shown high correlations with E2 and P4 in serum (Berthonneau et al., 1989; Gandara et al., 2007; Groschl et al., 2001). E2 concentrations were measured by direct assay using the DSL4800 Ultra-Sensitive Estradiol RIA ( Immunotech, Prague, Czech Republic), modified for saliva. P4 and testosterone concentrations were assessed by radioimmunoassay using Coat-A-Count kits (Siemens Healthcare Diagnostics, Deerfield, IL), which were adapted for saliva according to established laboratory protocols (Hampson & Kimura, 1988; Hampson & Morley, 2013a; Moffat & Hampson, 1996). Testosterone was measured in men only. Because of the pulsatile secretion of hormones, any

\(^2\) Since the assays of testosterone levels for about half of our male subjects were not finished, testosterone levels were not included in analyses in this report.
single measures of P4 or E2 may not perfectly reflect the average hormone levels or the menstrual phase. In addition, there are large inter-individual differences in E2 and P4 levels in women during their menstrual cycles (Sundström Poromaa et al., 2014). Therefore, only female participants in whom both P4 and E2 at either testing session were out of the normal range for that menstrual phase were excluded from further analyses. With the help of Dr. Hampson, the normal ranges of E2 and P4 were set to be: midluteal phase: P4>60pg/ml or E2>0.6pg/ml; early follicular phase: P4<35pg/ml or E2>1.1pg/ml.

2.5 COMT genotyping

All participants were genotyped for the COMT SNP rs4680 (G>A (nucleobase); Val\textsuperscript{158} Met). Genotyping was carried out in the Lab of Prof. Weihong Song in our Department, just upstairs from our Lab. Saliva samples were collected using Oragene kits (DNA Genotek, Ontario, Canada). Genomic DNA was extracted according to the protocol supplied by the manufacturer and then analyzed by PCR-restriction fragment length polymorphism (RFLP).

2.6 Subjective assessment of stress over the month preceding testing: The Perceived Life Stress Scale

The Perceived Stress Scale (PSS-10 version) questionnaire is a widely-used self-report psychological instrument for measuring how stressful a participant felt the previous month.
had been for him or her (Cohen et al., 1983). The PSS-10 version has 10 items, which are designed to tap how unpredictable, uncontrollable and/or overloaded participants might have found their lives over the past month. Each of the 10 items is rated on a 5-point Likert scale (0=never, 1=almost never, 2=sometimes, 3=fairly often, 4=very often). After inverting scores for the four positive items, a total score was computed by summing the scores of all items. The higher the score a participant had, the greater the perceived stress he or she indicated having experienced during the month before testing. The difference of the PSS score between two sessions for each participant was calculated. Such score differences were then averaged across all participants. Because we wanted to equate the baseline stress levels which participants had at the outset of their two testing sessions, participants were excluded whose difference in PSS scores, between the two sessions exceeded the upper or lower threshold of ± 2 standard deviations from the mean PSS score difference for all participants. As mentioned previously, 6 subjects (1 male and 5 females in the F-L group) were excluded from further analyses based on this criterion.

2.7 Subjective assessment of stress immediately before and after testing

In each testing session, participants filled out a questionnaire (see Appendix) before and after testing to report the degree of stress they were currently feeling along a 5-point scale. A space was provided on the questionnaire for participants to add anything that was not
captured by the questionnaire. We expected that at the end of the stress session, participants would report feeling more stressed (a) than at the end of the calm session and (b) than they were when they had arrived at the lab.

2.8 Cognitive tests

Participants took four computerized EF tests and one fluid intelligence test (which assesses reasoning, a higher order EF skill) in both sessions. Two different versions of these tests were used to minimize practice effects. In Session 1, ‘A’ versions of each test were administrated; in Session 2, ‘B’ versions were utilized. The first two EF tests assessed two of the three core EFs, inhibitory control and cognitive flexibility. They also required WM (the other core EF) but did not provide assays of that. To control for order effects within the session, the order in which the first two EF tasks were administered within a session was counterbalanced for each stress order and within each of the three subject groups\(^3\). The fluid intelligence test, Raven’s Advanced Progressive Matrices, was the third EF test to be administered, which was followed by a 10-min break to rest briefly, go to the restroom, walk

\(^3\)We decided to add the task order later after we have tested \(\frac{1}{4}~25\%\) of participants with Task Order 1 (taking the Flanker task as EF1 and the Hearts and Flowers (or Dots) task as EF2). To make the task order counterbalanced, for the remaining 75% of participants, relatively more participants were tested with the Task Order 2 (taking the Hearts and Flowers (or Dots) task as EF1 and the Flanker task as EF2)2 than Task Order 1.
around, stretch, get a drink, etc. The last two EF tasks (Re-Ordering Digital Span and Visuo-Spatial Working Memory Span), which assess WM, were introduced later. So we only have data on these 2 tests from around half (38 out of 70) of our participants.

2.8.1 Flanker/Reverse Flanker task:

In Block 1 of this Flanker/Reverse Flanker task, participants received the classic Flanker paradigm (Eriksen & Eriksen, 1974) where they were instructed to press a button on the left if the center stimulus was pointing left and a button on the right if the center stimulus was pointing right, ignoring the four flankers, two on either side of the center stimulus in Version A or the 8 flankers surrounding the center stimulus in Version B (Figure 3). This requires inhibiting attention to the distractors (the flankers). Block 2 presented a ‘Reverse Flanker’ condition (Munro et al., 2006), where participants now had to focus on the flankers (the outside stimuli) and ignore the center stimulus. In addition to requiring attentional control (inhibiting attention to the distractor [the center stimulus]), this also requires inhibiting the previously correct response (the response that had been correct during Block 1) and re-orienting where to focus one’s attention. Both blocks require remembering a rule; the memory requirements of the two blocks are comparable.
In Block 1, all stimuli were blue fish for Version A (or carets for Version B; see Figure 3). In Block 2, all stimuli were pink fish for Version A (or thick arrows for Version B; see Figure 3). In Block 3, participants were presented with classic Flanker trials (indicated by the

**Figure 3 Examples of stimuli used in our Flanker/Reverse Flanker task.**

Version A was used in Session 1 and Version B was used in Session 2. In Block 1, participants were tested on the classic Flanker paradigm where they were to press the button in the same direction in which the center stimulus was pointing, ignoring the flanking stimuli. All stimuli in this block were blue fish for Version A (or black carets for Version B). In Block 2, participants were tested on the Reverse Flanker paradigm, where they had to focus on the flanking stimuli and inhibit attending to the center stimulus. All stimuli in this block were pink fish for Version A (or thick black arrows for Version B).
stimuli being blue or carets) and Reverse Flanker trials (indicated by the stimuli being pink or thick arrows) pseudo-randomly intermixed (all participants received the same order of trials in Version A and a different order for version B that was the same for all participants).

In each of these three blocks, there were four subtypes of trials: 1) no-distractor trials (i.e., no flanking stimuli or no center stimulus); 2) congruent trials where the distractor(s) pointed in the same direction as the target; 3) incongruent trials where the distractor(s) pointed in the opposite direction from the target; and 4) neutral trials where the distractor(s) pointed in an irrelevant direction (either up or down; ‘irrelevant’ in that no response was associated with pointing in that direction). For all trials, stimuli appeared on the screen for 1500ms. The inter-trial interval lasted 500ms. The “Flanker Effect”, the RT difference between incongruent trials and easy (congruent and no-distractor) trials divided by the RT for the easy trials, has been demonstrated to be far larger in Block 3 (the mixed block) than in the classic Flanker or the Reverse Flanker block (Munro et al., 2006). The mixed block requires holding two sets of rules in mind: First participants must pay attention to the color or shape and use that to determine whether they should focus on the center stimulus or the flankers. Next participants must determine which way the target is pointing, and use that to determine on which side they should press. Hence, Block 3 places a heavy load on working memory. Block 3 also requires cognitive flexibility (participants must flexibly switch between the rules) and it requires inhibition of the non-relevant rule on each trial and inhibition of the distracting stimulus or stimuli on all trials.
Each of the first two blocks was preceded by instructions, a demonstration of the task, and a practice block where the participant received feedback on each trial. Before the practice block of Block 3, participants were reminded of the rules they had learned for the previous two blocks. Participants had to get at least three of the four trials in a practice set correct to demonstrate they understood the rule. If the participant made more than one error on the first practice set, the participant was reminded of the rules and given another set of 4 practice trials. No participant in the study failed to pass practice for any block. Each of the first blocks consisted of 17 trials (3 no-distractor trials, 5 congruent trials, 6 incongruent trials and 3 neutral trials [thus incongruent trials were 35% of all trials in each of those blocks])<sup>4</sup>. Block 3 contained 45 trials (22 Standard Flanker Trials and 23 Reverse Flanker trials; 8 no-distractor trials, 12 congruent trials and 16 incongruent trials, 9 neutral trials [these incongruent trials were 36% of all trials and switch trials (where the rule changed) were 55% of all trials in Block 3]).

The first trial in each block was excluded from RT analyses as RT on that trial would be unreliable because although participants knew the trials were starting, they did not know

<sup>4</sup>The numbers and percentages of different types of trials in the mixed block were the same between Version A and B. However, we did not find out that the percentages of different types of trials in the first two blocks were different between Version A and B until almost the end of this study. The first two blocks of Version B wrongly contained 2 no-distractor trials, 2 congruent trials, 10 incongruent trials and 3 neutral trials [thus incongruent trials were 63% of all trials in each of those blocks]. The effect of version was controlled for in all analyses of these tasks.
precisely when they would see the first stimulus. Trials with RTs < 250 ms\(^5\) were excluded for being too fast to have been in response to the stimuli (anticipatory responses; resulting in 4 trials being excluded). After these exclusions, accuracy (binary response (right or wrong) at the individual trial level) on valid responses and the percentage of correct responses (calculated by dividing the number of correct responses by the total number of valid responses) were used for further analyses. RTs that exceeded an upper or lower threshold of the mean RT for that participant ± 2 standard deviations were excluded from analyses for RT (for being outliers; resulting in 411 trials (4.7% of all valid responses) being excluded). RTs, as well as mean RTs for each block, on valid trials in that block on which he or she responded correctly were used for further analyses. That is, RTs on trials where subjects erred were not included in the RT calculations.

The indices we used as indicators of inhibitory control were: 1) Performance (i.e., accuracy and RT) on incongruent trials, and 2) The Flanker Effect:

\[
\frac{\text{Performance on incongruent trials} - \text{Performance on easy trials (congruent and no-distractor trials)}}{\text{Performance on easy trials (congruent and no-distractor trials)}}.
\]

We were also interested in the performance difference between inhibition in the context of switching rules and inhibition in a non-switching context. Hence, we calculated (a) RT on

\[\]

\(^5\) All further analyses for this task were repeated with two lower lowest valid RT to 200ms and 150ms to check whether the change of this RT limit produced different results or whether the limit of 250ms created an artificial flooring effect.
incongruent trials in Block 3 minus RT on incongruent trials in Blocks 1 and 2 and (b) the mean percent correct on incongruent trials in Blocks 1 and 2 combined minus the mean percent correct on incongruent trials in Block 3.

To measure cognitive flexibility, the indices we used were: 1) Performance in the mixed block (Block 3); 2) local switch cost (the performance difference between (a) switch trials (trials that required switching rules from the previous trial [either getting a Flanker trial after a reverse Flanker trial or getting a reverse Flanker trial after a Flanker trial]) and (b) non-switch trials); 3) same-side local switch cost (the performance difference between (a) switch trials where the rule changed from the previous trial but the site of the correct response did not change and (b) non-switch trials); 4) global switch cost-1 (the performance difference between all trials in Block 3 and all trials in the first two blocks); and 5) global switch cost-2 (the performance difference between only non-switch trials in Block 3 and all trials over Blocks 1 and 2 combined [Blocks 1 and 2 consisted only of non-switch trials]).

2.8.2 Hearts and Flowers (or Dots) task:

On this task (Davidson et al., 2006; Wright & Diamond, 2014), Block 1 is the congruent block, where participants were to press on the same side as the stimulus. The stimulus was either a red heart (Version A) or a black and white striped disc (Version B) and it was presented on the left or the right side of a horizontal rectangle (6cm x 18cm; see Figure 4). This demands little or no EFs, since our natural tendency is to activate the hand on the
same side as a stimulus. Block 2 is the incongruent block, where participants are to press on the side opposite the stimulus. The stimulus was either a red flower (Version A) or a grey disc (Version B). This requires inhibiting the natural tendency to activate the hand on the same side as a stimulus (Hommel, 2011; Lu & Proctor, 1995) and instead to activate the other hand. Wright & Diamond (2014) showed that it doesn’t matter whether the incongruent block is presented before or after the congruent one; hence the slower RT and greater percentage of errors of children on the incongruent block is due solely to the demand to inhibit pressing on the same side as the stimulus and not to having to inhibit the rule that was correct on Block 1 or to switch rules. Davidson et al. demonstrated using this task, as many others have shown using classic Simon tasks (Lu & Proctor, 1995) that when presented with an entire block of incongruent trials, adults do not show the Simon effect and are as fast and accurate on an incongruent block as on a congruent one.
In the mixed block (Block 3; where adults do show the Simon effect), congruent (heart or black-white striped disc) trials and incongruent (flower or gray disc) trials are randomly intermixed; participants must flexibly switch between the rules to press on the same side or the opposite side.
Blocks 1 and 2 each require holding one rule in mind and differ only in that Block 2 requires inhibitory control while Block 1 does not. Block 3 requires holding 2 rules in mind, flexibly switching between them, and inhibiting both the previous correct rule on switch trials and inhibiting the tendency to press on the same side as the stimulus on incongruent trials.

Hearts and Flowers differs from a Simon task (Diamond, 2013) in that Hearts and Flowers requires working memory (holding information in mind and manipulating it) whereas the Simon task requires only short-term memory (just holding information in mind). The rules on a Simon task are: for Stimulus A press on the right and for Stimulus B press on the left. So, if I tell you that Stimulus A will appear on the next trial, you know where to respond. However, for the Hearts and Flowers task, if I tell you Stimulus A will appear on the next trial, you do not yet know where to respond – because the correct response will require that you integrate object identity with object location. Knowing only the spatial location or only the identity of a stimulus is insufficient; you need to integrate the two. Moreover, we don’t have “same side” or “opposite side” hands. Therefore, to know where to respond requires translating the same or opposite rule into a right or left hand response. The Simon task requires no such mental translation.

For all trials in the Hearts and Flowers task, after a 500ms-fixation period with a crosshair presented in the center of the rectangle and another 500ms with only the rectangle
presented, stimuli appeared on the screen for 750ms. Each of the first two blocks was preceded by instructions, a demonstration of the task, and a practice block where the participant received feedback on each trial. Before the practice block of Block 3, participants were reminded of the rules they had learned for the previous two blocks. Participants had to get at least three of the four trials in a practice block correct to demonstrate they understood the rule. If a participant made more than one error on the first practice block, the participant was reminded of the rules and given another practice block. No participant in the study failed to pass practice for Block 1, 2 or 3.

There were 12 trials in each of the first two blocks and 33 trials (17 congruent trials and 16 incongruent trials) in the mixed block. Trials in each block were presented in the same pseudo-random order to each participant. All participants received the same order of trials in Version A and a different order for version B that was the same for all participants. The first trial in each block was excluded from further analyses. Trials with RTs < 250ms were excluded for being too fast to have been in response to the stimuli (resulted in 284 trials being excluded or 1.8% of trials). After these exclusions, accuracy (binary response (right or wrong) at the individual trial level) on valid responses and the percentage of correct responses (calculated by dividing the number of correct responses by the total number of

\[ \text{\ldots} \]

\[ \ldots \]

All further analyses for this task were repeated with two lower lowest valid RT to 200ms and 150ms to check whether the change of this RT limit produced different results or whether the limit of 250ms created an artificial flooring effect.
valid responses) were used for further analyses. RTs that exceeded an upper or lower threshold of the mean RT for that participant ± 2 standard deviations were excluded from analyses for RT (for being outliers; resulting in 145 trials (1.0% of all valid responses) being excluded). RTs, as well as mean RTs for each block, on valid trials in that block on which he or she responded correctly were used for further analyses. That is, RTs on trials where subjects erred were not included in the RT calculations.

The indices we used for assessing inhibitory control were: 1) performance on incongruent trials in Block 3 and 2) performance difference between incongruent trials and congruent trials in Block 3 (Blocks 1 and 2 are always subject to ceiling effects for adults). We were also interested in any difference in being able to exercise inhibition in the context of switching rules versus exercising inhibition in a non-switching context. Hence, we calculated (a) RT on incongruent trials in Block 3 (a switching context) minus RT on incongruent trials in Block 2 (a non-switching context) and (b) the mean percent correct on incongruent trials in Block 2 minus the mean percent correct on incongruent trials in Block 3. 

Our indices of cognitive flexibility were: 1) performance on all trials in the mixed block (Block 3); 2) local switch cost (the performance difference between switch and non-switch trials in Block 3); 3) same-side local switch cost (the performance difference between [a] switch trials where the response stayed on the same side as on the previous trial and [b] non-switch trials); 4) global switch cost-1 (the performance difference between all trials in Block 3 and all trials in the first two blocks); and 5) global switch cost-2 (the performance
difference between non-switch trials in Block 3 and all trials in the first two blocks [which were all non-switch trials]).

2.8.3 *Raven’s Advanced Progressive Matrices (RAPM):*

All versions of Raven’s Matrices are widely-used, highly-regarded tests of non-verbal logical reasoning (also known as fluid intelligence (Perfetti et al., 2009; Raven et al., 1998; Unterrainer et al., 2006)). Not surprisingly, performance on this test correlates extremely highly with other EF measures (Duncan et al., 1996; Kane & Engle, 2002). As per John Raven’s recommendation (personal communication), the even-numbered trials were presented in Session 1 and the odd-numbered trials were presented in Session 2. (Unfortunately this resulted in the Session 2 version being a bit easier, since difficulty increases over trials; thus Trial 1 is easier than Trial 2, Trial 11 is easier than Trial 12, etc. Therefore, we converted total scores into z-scores for each version separately to control for the version effect). Participants were to identify the missing component in a series of figural patterns (see Figure 5). Two practice trials were utilized to reinforce and demonstrate understanding of the test. After each practice trial, the experimenter pointed out the rules that govern the changes within the matrix to explain why the participant’s answer was correct (or incorrect). No participant failed to demonstrate understanding of the rules. Participants were given x minutes to complete 18 trials in each session. A total score was computed by summing the number of items correctly answered.
Re-Ordering Digit Span task:

Classic digit span tasks require repeating back digits in the same or reverse order in which they were heard (Wechsler, 1981). Forward digit span only requires keeping information in mind, not manipulating it in any way, and so is a measure of short-term memory, not WM (Diamond, 2013). Backward digit span requires manipulating the information (reversing the order) unless one circumvents that by building a mental image of one number as the digits are heard (e.g., 4719), and then simply looking at that number and...
‘reading off’ the digits it contains from back to front. Re-Ordering Digit Span, where participants are to repeat the digits they just heard in numerical order (e.g., after hearing 7, 2, 10, repeating back 2, 7, 10), cannot be done without mentally manipulating the information; thus it is clearly a measure of WM. Three example trials were given for teaching purposes and three practice trials were provided to determine understanding of the rule.

Starting from a sequence of three digits, an additional digit was progressively added to the length of the sequence after a participant got two trials correct in a row at a single level. If the participant missed two trials in a row, the length of the sequence was reduced by one digit. If the pattern of performance over 4 trails of the same sequence length was alternating correct and error (i.e. “correct, error, correct, error”), the participant was also considered to have failed at that level, and the length of the sequence was reduced by one digit. If a participant failed a given level twice, the task ended after he or she got two trials correct in a row at the next easier level. The task also ended when a participant got two trials correct in a row at our longest sequence length (7 digits). A participant’s digit span was considered to be the maximum length at which that participant got two trials in a row correct. Self-corrections, if they occurred, were recorded but not counted as a correct response. Number sequences only included the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, and 20. Note that each is one syllable only. The provisos which were used to prevent the task from being too easy included: 1) for two digits: the second digit was never one greater or one less than the first; 2) for three
digits or above: i) the sequence was never a sequence of three consecutively numbered digits; ii) the sequence was never a sequence of three odd (or three even) consecutively numbered digits; iii) over all the trials, the initial digit value was distributed throughout the available digits, so that sequences did not always begin with a low, or middle, or high value. The interface that was used by the experimenter for recording was presented in Figure 6.

Figure 6 The computer interface that was used by experimenter for recording in the Re-Ordering Digit Span task.

Participants were to repeat the digits they just heard in numerical order.
2.8.5 Visuo-Spatial Working Memory Span task:

This task is a computerized variant of the Corsi Block test (Chechlacz et al., 2014; Corsi, 1972; Lezak, 1983; Miyake et al., 2000) with a masking stimulus (Breitmeyer & Ogmen, 2000) between presentation of the sequence and when a participant could respond on each trial. The traditional Corsi Block or Spatial Span task (Alloway, 2007; Alloway et al., 2009) might be considered a measure of short-term memory (rather than WM) because subjects are not required to manipulate the stimuli in any way. When a masking stimulus is inserted, however, the task then confirms to the definition of WM that Kane and Engle (2002) have long championed; i.e., holding information in mind and exercising interference control (blocking or inhibiting other information from interfering with what you are trying to hold in mind).

Our spatial span task was administrated through the use of a touch-screen monitor. Participants were presented with 16 yellow squares in a 4 x 4 grid and watched a sequence of squares in that grid turn black one at a time, each for 500 ms, and then return to yellow before the next square turned black (see Figure 7). After the sequence, the backward masking stimulus was presented which consisted of the four squares making up a diagonal from the upper left to lower right turning black for 500 ms. Then participants were asked to tap the squares in the same order in which the squares had turned black. For each square that was correctly tapped, a small picture of a cute animal appeared in that square for 500ms. If a square was incorrectly pressed, the square would turn red, and the correct sequence would be
presented again and the program would move on to the next trial. Two example trials for teaching purposes and two practice trials were given to ensure understanding of the rule. The provisos which were used to prevent the task from being too easy included: 1) the pattern for each trial was never confined to a sub-portion of the grid, and consecutive squares in a trial were never adjacent (top, down, left or right); 2) for three boxes, i) no two boxes (of the three) were in the same row or same column; ii) none of three boxes were adjacent, or even diagonally touching; 3) for four or more boxes, i) no pattern was confined to a sub-portion of the matrix; ii) none of boxes were adjacent but touching diagonally was OK; 4) for two consecutive trials, the first box in a sequence did not begin in the same side (i.e. top, bottom, left or right); 5) over all the trials, the first box in a sequence was distributed equally in all four sides in the grid and did not always begin in an outer row or column but also began in an inner row or column.

Starting from a sequence of three (same as for the Re-Ordering Digit Span task), an additional square was added to the length of the sequence after a participant got two trials correct in a row at a single level. If the participant missed two trials in a row, the length of the sequence was reduced by one. If a participant failed a given level twice, the task ended after that participant got two trials correct in a row at the next lower level. The task also ended when a participant got two trials correct in a row at a sequence length of 7, our longest sequence. Spatial working memory span was taken to be the maximum length at which the participant got two trials in a row correct.
2.9 Statistical analyses

To check whether the normality assumption for parametric tests was met, the Shapiro-Wilk test for original raw data and residuals was performed and interpreted in conjunction
with histograms, Q-Q plots and values of skew and kurtosis. For the assumption of
homogeneity of variance, Levene’s test was used. Non-parametric analyses were performed
if the normality assumption or the assumption of homogeneity of variance was not met.
Stress responses (e.g. cortisol, the ratio of cortisol to DHEA, BP (both systolic and diastolic
BP) and HR) were analyzed by separate mixed-design analyses of variance (ANOVAs) with
time as a repeated-measures factor, condition (more stressful versus calmer) as a within-
subject factor, and subject-group (M, F-L and F-H) and stress order as between-subject
factors (Field, 2009). When there was a significant difference among groups, Bonferroni-
Holm-corrected between-subject pairwise comparisons using t-tests were performed to see
which two groups differed significantly from one another (Holm, 1979). In addition,
Bonferroni-Holm-corrected within-subject pairwise comparisons using t-tests were
performed to investigate differences in stress levels between the two conditions (more
stressful vs calmer) at each time-point. As noted previously, the last two tasks were added
later, so we only have the last two measures of BP and HR and the last sample for cortisol
levels for some of our participants. Therefore, the last two measures of BP and HR and the
last measure of cortisol levels were not included in the corresponding ANOVAs but included
in separate within-subject t-tests.

For EF measures, a log transformation was used before analyzing RT of the first two
tasks to reduce significant skew. As mentioned previously, we found significant differences
between Version A and Version B in RTs on the Flanker/Reverse Flanker task and in the
total scores of RAPM. Thus, to control for the effects of differences between versions, we used 1) the version number as a fixed factor in the linear mixed-effects models or generalized linear mixed model (described below) or 2) entered stress order*condition into the mixed-design ANOVAs (the version effect was included as part of order/carryover effects). In addition, for RAPM, we converted total scores into z-scores for each version separately. To take advantage of information from each trial of a task (trial nested within trial-block or task), the transformed RT was analyzed by a linear mixed-effects (LME) model with condition (more stressful versus calmer), subject group (M, F-L, F-H), and their interaction as fixed-effect variables and with SubjectID and SubjectID by session as random-effect variables (Field, 2009). Age and the PSS score were included as covariates, and stress order, task order, and version of the task (A or B) were included as fixed factors to control for their effects. At the individual trial level, accuracy is binary (correct or incorrect). With the same fixed- and random-effect variables, accuracy (right or wrong) on each of the first two tasks was analyzed using a generalized linear mixed model (GLMM; (Mehta & Patel, 1995). For other indices that were calculated using mean RT (e.g. the Flanker Effect), % correct in the first two tasks, z-scores of performance in RAPM and longest spans in the last two WM tasks were analyzed by separate 2(Condition) by 3 (Subject group) ANOVAs controlling for the order effects and the effects of age and difference in the PSS score between the stressful versus calmer conditions. To better illustrate the subject group by condition interaction term which we were interested in, univariate ANOVAs were used for the difference scores.
between calm and stress conditions on indices (e.g. Local Switch Cost) for the first two tasks with subject group as the fixed factor. These difference scores were calculated by subtracting the index in the stress condition from the index in the calm condition. As smaller values for these indices indicate better EFs, positive values for these difference scores would indicate stress-induced improvements in EF performance, whereas negative values would indicate stress-induced impairments in EF performance. Moreover, composite scores for difference scores in inhibitory control and in cognitive flexibility were also analyzed. In addition, for each EF task, each subject group was further divided into two subgroups based on their EF performance during calm condition (well-performing subgroup had performance better than or equal to the median of the subject group; poorly-performing group had performance worse than the median of the subject group). Independent t-tests were performed on the difference in their performance between two conditions (more stressful and calmer) between well-performing subgroup vs. poorly-performing subgroup to determine whether stress differentially affected their performance. To correct for conducting multiple pairwise comparisons, the Bonferroni-Holm correction was used. Greenhouse-Geisser corrected p-values were used when the assumption of sphericity was violated. The alpha level was set at \( \leq 0.05 \). All analyses were conducted using SPSS 23.0 (IBM, Chicago, USA).
3. Results

3.1 Demographic characteristics

The three groups did not differ in age, education levels or ethnic composition (see Table 1).

<table>
<thead>
<tr>
<th></th>
<th>Men</th>
<th>Women (early follicular [low E2])</th>
<th>Women (midluteal [high E2])</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of subjects</td>
<td>26</td>
<td>20</td>
<td>24</td>
</tr>
<tr>
<td>Age (mean±SD)</td>
<td>24.5±4.9</td>
<td>23.4±4.0</td>
<td>24±4.0</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>65.4% Caucasian; 7.7% Middle Eastern; 7.7% South Asian; 3.8%; Hispanic; 15.4% East Asian</td>
<td>55% Caucasian; 10% Middle Eastern; 15% South Asian; 5%; Multiracial; 15% East Asian</td>
<td>58.3% Caucasian; 4.2% Middle Eastern; 4.2% South Asian; 12.5% Hispanic; 16.7% East Asian; 4.2% Native American</td>
</tr>
<tr>
<td>Estradiol (pg/ml) (mean±SD)</td>
<td>0.84±0.46</td>
<td>0.81±0.43</td>
<td>1.13±0.51</td>
</tr>
<tr>
<td>Progesterone (pg/ml) (mean±SD)</td>
<td>24.0±28.0</td>
<td>27.4±20.6</td>
<td>104±86.5</td>
</tr>
</tbody>
</table>

Table 1 Characteristics of participants in the three subject groups.

3.2 Sex hormone assessments

As expected, independent-sample t-tests showed that, compared to the F-L or M groups, the F-H group had significantly higher levels of E2 and of P4 (for F-H vs F-L: E2: $t(39) = 2.50$, $p = 0.017$; P4: $t(28.9) = 5.15$, $p < 0.001$ (Levene’s test indicated unequal variances ($F = 13.1$, $p = 0.001$) so degrees of freedom were adjusted from 39 to 28.9); for F-H vs. M: E2: $t$
(47) = 2.32, \ p < 0.02; P4: \ t (29.3) = 5.47, \ p < 0.001 \ (\text{Levene’s test indicated unequal variances } (F = 13.0, \ p = 0.001) \ so \ degrees \ of \ freedom \ were \ adjusted \ from \ 47 \ to \ 29.3); \ see Table 1 \ and \ Figure 8A \ and \ 8B). \ The \ average \ E2 \ level \ in \ the \ M \ group \ was \ slightly \ but \ not significantly \ higher \ than \ average \ E2 \ levels \ in \ the \ F-L \ group. \ Average \ P4 \ level \ in \ our \ male group \ was \ also \ comparable \ to \ those \ of \ the \ F-L \ group.
Figure 8 Mean estradiol (A) and progesterone (B) levels.

E2 and P4 levels were measured from saliva samples collected in each testing session immediately before testing. As expected, the F-H group had significantly higher E2 and P4 levels than the F-L and M groups.

M = men. F-L = women tested during early follicular phase (low E2 and P4 levels). F-H = women tested during midluteal phase (high E2 and P4 levels). Error bars are standard errors. *p < 0.05, ***p < 0.001.
3.3 Stress levels of participants

Did we succeed in stressing participants in all subject groups during the session intended to be stressful? Were participants in all groups more stressed during that session than during the session intended to be un-stressful?

3.3.1 Blood pressure and heart rate:

Separate 2 (condition: more stressful versus calmer) by 4 (time: baseline, after EF1, after EF2, after EF3) repeated-measures ANOVAs for absolute values of systolic BP, diastolic BP and HR at each time-point, and Bonferroni-Holm-corrected within-subject t-tests comparing the stressful versus calmer conditions, were performed.

Within-subject t-tests indicated that the baseline systolic BP levels were significantly higher at the start of the stress session compared with the start of the calm session when all three groups were combined \( (t(69) = 2.371, p = 0.02) \) and for the F-L group alone \( (t(19) = 2.75, p = 0.01) \) but not for the other two groups. Baseline BP levels were significantly higher in the M group compared to the other two groups for both conditions \( (M \text{ vs. F-L: } t(44) = 3.74, p = 0.001; M \text{ vs. F-H: } t(48) = 5.18 \ p < 0.001) \), but the two female groups did not differ from each other for conditions.

The ANOVA looking at **absolute levels of systolic BP** yielded significant main effects for condition \( (F(1, 66) = 52.2, p < 0.001, \eta^2 = 0.44) \) and for subject group \( (F(2, 66) = 19.4, p < 0.001, \eta^2 = 0.37) \). There was also a marginally significant time by condition interaction
The interactions of condition by subject-group and condition by time by subject-group did not reach significance, indicating that there was no significant sex/group difference in the effects of our stress induction procedure on systolic BP. To illustrate this graphically, the differences in absolute systolic BP levels between the more stressful versus calmer conditions for each subject group are presented in Figure 10A. Follow-up simple contrasts comparing systolic BP levels later in the session to baseline levels revealed that, compared to the calm condition, participants in the stress condition showed greater increases in systolic BP after EF 1 ($F(1,66) = 5.14, p = 0.027, \eta^2 = 0.07$) and the systolic BP remained high during the rest of stress condition (after EF2: $F(1,66) = 3.96, p = 0.05, \eta^2 = 0.06$; after EF3: $F(1,66) = 5.60, p = 0.02, \eta^2 = 0.09$). Bonferroni-Holm-corrected within-subject t-tests confirmed this interaction and indicated that systolic BP was significantly higher in the stress condition after each EF tasks for all subjects compared to systolic BP in the calm condition (after EF1: $t(69) = 5.34, p < 0.001$; after EF2: $t(69) = 5.06, p < 0.001$; after EF3: $t(69) = 4.76, p < 0.001$; after EF4&5: $t(69) = 2.32, p = 0.02$). The same was true when looking at just the F-L group alone (after EF1: $t(19) = 2.75, p = 0.01$; after EF2: $t(19) = 2.38, p = 0.02$; after EF3: $t(19) = 3.29, p = 0.004$; after EF4&5: $t(11) = 3.28, p = 0.007$). This was also true after EF1, EF2 and EF3 for the M group (after EF1: $t(25) = 3.23, p = 0.003$; after EF2: $t(25) = 2.81, p = 0.01$; after EF3: $t(25) = 4.28, p < 0.001$) and after EF1 and after EF2 for the F-H group (after EF1: $t(23) = 3.14, p = 0.005$; after EF2: $t(23) = 3.57, p = 0.002$). Post hoc tests with Bonferroni-Holm-corrections
revealed that men had significantly higher BP than women (M vs. F-L: \( t(44) = 4.51, p < 0.001 \); M vs. F-H: \( t(48) = 5.91, p < 0.001 \)).
(B)
Figure 9 Results for autonomic indicators of stress.

Presented here are the mean levels of systolic blood pressure (A), diastolic blood pressure (B) and heart rate (C) for each time-point at which they were measured for the more stressful and calmer conditions (white and grey bars, respectively) for each subject group and for all participants combined. M = men. F-L = women tested during early follicular phase (low E2 and P4 levels). F-H = women tested during midluteal phase (high E2 and P4 levels). Error bars indicate standard error. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. 
Figure 10 Differences in autonomic indicators of stress between the two conditions (more stressful and calmer) for each subject group.

Presented here are the mean levels of differences in systolic blood pressure (A), diastolic blood pressure (B) and heart rate (C) between the more stressful versus calmer conditions for each time-point at which they were measured for each subject group (indicated by different colors). M = men. F-L = women tested during early follicular phase (low E2 and P4 levels). F-H = women tested during midluteal phase (high E2 and P4 levels). F-L&F-H = all women. Error bars indicate standard error.
No significant differences were found for diastolic BP at baseline between the more stressful versus calmer conditions for any subject group alone or all combined. Among three subject groups, the only significant difference found at the baseline was between the M group and the F-L group in the stress condition with significantly higher diastolic BP in the M group than in the F-L group (\( t(48) = 2.43, p = 0.05 \)).

The ANOVA for absolute levels of diastolic BP indicated significant main effects of condition (\( F(1,66) = 11.2, p = 0.001, \eta^2 = 0.15 \)) and of time (\( F(3,198) = 3.50, p = 0.02, \eta^2 = 0.05 \)), a marginally significant effect of subject group (\( F(2,66) = 2.70, p = 0.08, \eta^2 = 0.08 \)) and a significant interaction of time by condition (\( F(3,198) = 4.30, p = 0.006, \eta^2 = 0.06 \)). No significant effects of condition by subject-group or condition by time by subject-group were found (see Figure 10B), which indicates there was no significant sex difference in the effects of our stress induction procedure on diastolic BP. Follow-up simple contrasts comparing diastolic BP levels later in the session to baseline levels showed that, compared to the calm condition, participants in the stress condition showed larger increases in diastolic BP beginning after EF 2 (\( F(1,66) = 9.27, p = 0.003, \eta^2 = 0.12 \)) and continuing through to after EF3 (\( F(1,66) = 8.53, p = 0.005, \eta^2 = 0.11 \)). Bonferroni-Holm-corrected within-subject t-tests revealed that diastolic BP was significantly higher in the stress condition than in the calm condition after EF2 and after EF3 for all subjects (after EF2: \( t(69) = 3.81, p < 0.001 \); after EF3: \( t(69) = 3.43, p = 0.001 \)), after EF3 for men (\( t(25) = 3.42, p = 0.006 \)), after EF2
and EF3 for women tested when their E2 levels were low ($p < 0.045$), and after EF2 for women tested when their E2 levels were high ($t(23) = 3.14, p = 0.015$; see Figure 9B).

For HR, no significant differences were found at baseline between the more stressful versus calmer conditions for any subject group alone or all combined, or among the three subject groups for either condition. The mixed-design ANOVA for absolute HR yielded a significant main effect of time ($F(2.61, 172) = 8.25, p < 0.001, \eta^2 = 0.11$) and a significant time by condition interaction ($F(2.36, 156) = 3.28, p = 0.03, \eta^2 = 0.05$; see Figure 9C). No significant effects of condition by subject-group or condition by time by subject-group were found (see Figure 10C), which indicates there was no significant sex difference in the effects of our stress induction procedure on HR. Follow-up simple contrasts related to HR after EF1 showed that, compared to the calm condition, participants in the stress condition showed stronger increases in HR after EF3 ($F(1,66) = 8.04, p = 0.006, \eta^2 = 0.11$). Bonferroni-Holm-corrected within-subject t-tests revealed that participants’ HRs were significantly higher in the stress condition than in the calm condition after EF3 for all subjects ($t(69) = 2.60, p = 0.033$), and for the F-H group alone ($t(23) = 3.21, p = 0.012$).

**3.3.2 Cortisol and cortisol/DHEA ratio:**

There were several outliers detected in the histograms for baseline cortisol levels when checking for normality, which inflated the means and standard deviations. Therefore, cortisol levels were screened to eliminate outliers from each subject group prior to analyses. Outliers
were defined as having baseline cortisol levels outside a 1.5 Inter-Quartil-Range above the upper, or below the lower, quartile for each subject group. Cortisol levels and cortisol: DHEA ratios from participants who had baseline cortisol levels that qualified as outliers were excluded from analyses. Specifically, data from 8 participants (2 men and 3 women in each female group) were excluded. There were no significant differences in baseline cortisol levels between the two conditions (more stressful and calmer) for all subjects combined or for each subject group alone. There were also no significant differences in baseline cortisol levels among subject groups for either the more stressful or calmer condition. A 2 (condition) by 4 (time: baseline, after EF2, after EF3, and after a 10-min break) repeated-measures ANOVA for absolute cortisol levels revealed a significant main effect of time ($F(1.80, 91.6 = 5.46, p = 0.007, \eta^2 = 0.097$) and a significant time by condition interaction ($F(2.01, 102) = 4.47, p = 0.01, \eta^2 = 0.08$). Follow-up simple contrasts comparing cortisol levels later in the session to cortisol levels at baseline showed that, compared to the calm condition, participants in the stress condition showed larger increases in cortisol after EF2 ($F(1.52) = 5.08, p = 0.03, \eta^2 = 0.09$) and after EF3 ($F(1.52) = 6.27, p = 0.02, \eta^2 = 0.11$). Furthermore, there was a significant condition by time by subject-group interaction ($F(4.02, 13.34) = 2.65, p = 0.04, \eta^2 = 0.09$) which indicates that the condition by time interaction described previously was different among subject groups (see Figure 12A). Follow-up simple contrasts comparing the two conditions showed that there was a significant difference among subject groups for cortisol levels after EF3 compared to baseline levels (the F-L group showed a
stronger increase than the F-H group, $F(2, 52) = 4.17, p = 0.02, \eta^2 = 0.14$). Bonferroni-corrected within-subject t-tests indicated that for all subjects combined and both sessions combined, compared to baseline levels, there was an overall decrease in cortisol levels during testing (compared to baseline: after EF1: $t(69) = 4.49, p < 0.001$; after EF2: $t(69) = 3.18, p = 0.003$; after EF3: $t(69) = 2.03, p = 0.048$). Cortisol levels were marginally significantly higher 10 mins after EF3 in the stress condition than in the calm condition for all subject groups combined ($t(61) = 1.73, p = 0.09$). No other even marginally significant results were found from Bonferroni-Holm corrected within-subject t-tests comparing cortisol levels in the more stressful versus calmer conditions at any other time-point for any subject group. As you can see in Figure 11A, however, there was a trend in each and every subject group for participants’ cortisol levels to be higher in the stress condition, especially beginning 20-45 min after the onset of testing (after EF2 and after EF3).

For cortisol/DHEA ratio, there were no significant differences in baseline levels between the two conditions (more stressful and calmer) for all subjects combined or the M group or the F-H alone, but a marginally significant difference for the F-L group (higher at the start of the calm condition; $t(16) = 2.06, p = 0.06$). Marginally significantly higher cortisol levels at baseline were found in the M group compared to the F-L group in the stress condition ($t(39) = 2.36, p = 0.07$). No significant main effects or interactions were found from the ANOVA for the absolute values of this ratio (see Figures 11B & 12B).
(A)
Figure 11 Results for HPA-axis indicators of stress.

Presented here are the mean levels of cortisol (A) and cortisol/DHEA ratio (B) for each time-point at which they were measured during the more stressful and calmer conditions (red and blue lines, respectively) for each subject group. M=men. F-L=women tested during early follicular phase (low E2 and P4 levels). F-H=women tested during midluteal phase (high E2 and P4 levels). Error bars indicate standard errors.
3.3.3 Subjective assessments of stress:

Since this outcome variable is ordinal, the Wilcoxon signed-rank test was used for within-subject comparisons between the two conditions, as well as for pre-test versus post-test comparisons for each condition. There were no differences in baseline levels of subjective stress at the start of the stress session versus the start of the calm session for all subjects combined or for any subject group alone. However, as expected, significantly higher subjective levels of stress were reported from all subjects combined after the stress session than after the calm session ($Z = 2.920, p < 0.004$) and from the M group and the F-H group ($Z = -0.934, p = 0.05; Z = -1.947, p = 0.05$, respectively) but not from the F-L group. Subjects reported feeling more stressed after testing than before in both the calm and the stress sessions for all subjects and for each subject group individually (more stressful condition: M: $Z = -2.76, p = 0.006$; F-L: $Z = -2.997, p = 0.003$; F-H: $Z = -3.036, p = 0.002$; for calmer
condition: M: \( Z = -2.14, p = 0.032 \); F-L: \( Z = -3.17, p = 0.002 \); F-H: \( Z = -2.02, p = 0.043 \); see Figure 13). But changes in self-reported stress levels were greater in the stress condition than in the calm condition for all subjects (\( Z = -2.35, p = 0.02 \)).

To investigate possible sex difference in absolute self-reported levels of stress or in differences in self-reported stress levels between the two conditions, the Kruskal-Wallis and Mann-Whitney tests were used. They revealed no significant differences 1) in the difference between the two conditions (stress of calm) for pre-test stress levels or post-test stress levels across subject groups or between men versus women or 2) in changes from pre-test to post-test in each condition across the subject groups or between men versus women, which indicates there were no significant sex differences in the effects of our stress induction procedure on subjective levels of stress (see Figure 14).
Figure 13 Mean self-reported stress levels throughout the more stressful and calmer conditions for each sex group.

Significantly higher stress levels were reported after testing session in the stress condition than in the calm condition. M=men. F-L=women tested during early follicular phase (low E2 and P4 levels). F-H=women tested during midluteal phase (high E2 and P4 levels). Results were analyzed by Bonferroni-Holm-corrected paired sample t-tests. Error bars are standard error. * p<0.05, ** p<0.01.
No significant sex difference was found in differences in self-reported stress levels comparing the stress and calm conditions. M=men. F-L=women tested during early follicular phase (low E2 and P4 levels). F-H=women tested during midluteal phase (high E2 and P4 levels). Results were analyzed by Bonferroni-Holm-corrected paired sample t-tests. Error bars are standard error.
3.4 Executive function (EF) cognitive tests

3.4.1 Order effect:

We included a condition by session variable in each LME for RT and each GLMM for accuracy for the first two tasks, and condition X stress order variable in mixed-design ANOVAs for RAPM and two WM tasks to check for order effects. No significant effect of this interaction was found in any of these analyses.

3.4.2 Flanker/Reverse Flanker task:

For indices assessing inhibitory control using this task: The analyses (LME on log-transformed RT and GLMM for accuracy) for incongruent trials across all 3 blocks of trials, all trials in Block 2, or all incongruent trials just in Block 3, yielded no significant differences between the calm and stress sessions. For indices (e.g. the Flanker Effect) which were calculated using mean RTs, mixed-design ANOVAs did not find any significant effects of condition or interactions. For indices that were calculated using % correct, as the normality assumption was not met, non-parametric tests were used. Wilcoxon signed-rank tests showed that women tested when their E2 levels were high had a significantly larger value in the index (difference in percentage of correct responses on incongruent trials versus easy trials (congruent and no-distractor trials) in Block 2 and 3 combined was divided by percentage of correct response on easy trials) when stressed than when calm ($Z = 2.02, p = 0.044$), whereas the other two groups did not show a significant difference in this index
between the more stressful and calmer conditions (see Figure 15A). This indicates that on this one measure, stress significantly impaired the inhibitory control of the F-H group but it did not affect the inhibitory control of the other two groups.

The Mann-Whitney U test was then used to look at differences among groups on the difference scores between the calm versus stress conditions (stress-induced improvements or impairments) on an the above index. A significant difference was found between men and women tested when their E2 levels were high \((U = 172, z = 1.96, p = 0.050)\)\(^7\). When combining the data of the M group with data of the F-L group to compare with the data of the F-H group, that difference too, was significant \((U = 301, z = -2.03, p = 0.043; \text{ see Figure 15B})\). For indices as well as difference scores and composite scores that were calculated by mean RTs, neither the effect of subject-group in either condition nor group differences in the effect of condition reached significance \((F_{s} < 0.93, p_{s} > 0.40)\). However, the performance of men always tended to show a stress-induced improvement (positive values for difference scores or composite scores) while the performance of women tested when their E2 levels were high always showed a stress-induced impairment (negative values). Women tested

\(^7\) When using the more robust univariate ANOVA, there was a marginally significant main effect of subject-group \((F(2, 63) = 3.25, p=0.046)\). Bonferroni-corrected pairwise comparisons indicated the M group had a significantly higher mean difference score than the F-H group \((t(44) = 2.52, p = 0.015)\). The F-L group had a marginally significantly higher mean difference score than the F-H group \((t(36) = 1.74, p = 0.090)\).
when their E2 levels were low showed a profile intermediate between the two (e.g Figure 16C).

After outliers were removed, women tested when E2 levels were high still had a significantly larger difference in their percentage of correct responses on incongruent trials versus easy trials (congruent and no-distractor trials) in Block 2 and Block 3 when stressed than when calm ($Z = 2.48, p = 0.013$). The differences between two conditions remained non-significant for the other two groups ($ps > 0.5$)
After removing outliers, the between-group comparison became marginally significant (U=300.5 Z=1.77, P=0.078)
For indices assessing cognitive flexibility, analyses using the LME model for log-transformed RT on switch trials revealed a significant main effect of condition ($ F (1,60) = 5.23, p = 0.03$) with significantly shorter RTs in the stress condition than in the calm condition, however, there was no significant condition by subject-group interaction. No significant results were found for accuracy from the GLMM. The ANOVA for same-side local switch cost (RT difference between switch trials where response size did not change from the preceding trials and non-switch trials) yielded a marginally significant main effect.
of condition \((F (1, 57) = 3.52, p = 0.07)\) with smaller local same-side switch costs in the stress condition than in the calm condition.

Next, global switch costs were compared in the more stressful versus calmer conditions. It did not matter what variable was used to assess global switch costs – (a) the RT difference between all trials in Block 3 and all trials in the first two blocks (global switch cost-1), (b) the RT difference between only non-switch trials in Block 3 and all trials in the first two blocks since the first two blocks (Blocks 1 and 2) consisted only of non-switch trials (global switch cost-2; see Figure 16 A-B); (c) the difference in accuracy between all trials in Block 3 and all trials in the first two blocks, or (d) the difference in accuracy between non-switch trials in Block 3 and all trials in the first two blocks – men showed a stress-induced improvement (positive values for difference scores in means of indices which were calculated by RT or mean ranks of indices which were calculated by % correct) for global switch costs while women tested when their E2 levels were high showed a stress-induced impairment (negative values) in global switch costs. Women tested when their E2 levels were low showed a profile intermediate between the M and F-H groups. The same was true for composite scores of difference scores assessing inhibitory control and/or cognitive flexibility (see Figures 16C & D). However, neither the effect of subject-group nor group differences in the differences between the two conditions reached significance (for indices
calculated by RTs: both \( F_s < 1.27, ps > 0.27 \); for indices calculated by % correct: \( ps > 0.35 \)^{10}.

No significant difference in the effects of stress on these indices was found between well-performing groups vs. poorly-performing groups with any subject group. However the well-performing group always showed a tendency toward more impairment from stress than the poorly-performing group.

HR after EF2 relative to baseline was positively correlated with the Flanker effect for all trials for the F-L group (\( r=0.371, p = 0.028 \)), which indicates the larger the HR response was, the worse the performance of those in the F-L group. No other statistically significant correlation was found between any performance variable and any physiological indicator of stress (HR, cortisol) relative to baseline levels.

---

^{10} After lowering the lowest valid RT to 200ms or 150ms, analyses on group difference yielded similar results.
Stress facilitated performance of men but impaired performance of women with high E2, however, no significant main effect of subject group or significant group differences were found. The composite scores for inhibitory control consist of 1) the RT difference between incongruent trials versus easy trials in Block 2 and 3 combined and 2) the RT difference between incongruent trials in Block 3 versus incongruent trials in Blocks 1 and 2. The composite scores for inhibitory control consist of 1) Local Switch Cost in Block 3 and 2) Global Switch Cost-2. M=men. F-L=women tested during early follicular phase (low E2 and P4 levels). F-H=women tested during midluteal phase (high E2 and P4 levels). Mean and error bars were based on adjusted statistics. Error bars indicate standard error.

3.4.3 Heart and Flower (or Dots) task:

Results for this task were analyzed using the same statistical methods as described for the Flanker/Reverse Flanker task. No significant condition by subject-group interaction was found in the LME for RT, GLMM for accuracy, or mixed-design ANOVAs. For the difference scores between the stress condition and the calm condition on 1) indices assessing inhibitory control: a) RT difference between incongruent trials and congruent trials in Block
3 and b) RT difference between incongruent trials in switching environment (Block 3) and incongruent trials in steady state (Block 2), and 2) indices assessing cognitive flexibility: a) local switch cost in Block 3 (RT difference between switch trials and non-switch trials in Block 3 ) and b) global switch cost-1(RT difference between all trials in Block 3 and all trials in the first two blocks ), men consistently showed a tendency toward stress-induced improvements, women tested when their E2 levels were high consistently showed a tendency toward stress-induced impairments, and women tested when their E2 levels were low tended to perform at a level intermediate between the M and F-H groups (see Figure 17 A-D). The same was true for composite scores of difference scores for inhibitory control and cognitive flexibility (see Figures 17E & F) and composite scores across the Flanker/Reverse Flanker task and the Heart and Flower (or Dots) task (see Figure 17G & H). However, the data analyses found no statistically significant condition by subject-group interaction or significant differences among subject groups in the difference in performances between the more stressful versus calmer conditions on this task\(^{11}\). The lack of a significant group difference in analyses on accuracy in this task might be due to the ceiling effects on accuracy (more than 90% of participants got 100% for Block 1 or Block 2 and around 25% of participants got 100% for Block 3).

\(^{11}\) After lowering the lowest valid RT to 200ms or 150ms, analyses on group difference yielded similar results.
No significant difference in the effects of stress on these variables was found between well-performing groups vs. poorly-performing groups for any subject group (all \( ps > 0.21 \)).

HR after EF2 relative to baseline was positively correlated with the Flanker effect across all trials for the F-H group \((r = 0.326, p = 0.043)\), which indicates the larger the HR response, the worse the performance of those in the F-H group. No other statistically significant correlation was found between any other performance measure for this task and any physiological indicators of stress (HR, cortisol) relative to baseline levels \((ps > 0.11)\).
Figure 17 Mean difference scores for RT difference between incongruent trials and congruent trials in Block 3 (Panel A) and RT difference between incongruent trials in a switching environment (Block 3) and incongruent trials in an environment where the rule did not change (Block 2; Panel B), Local Switch Cost for RT in Block; Panel C), and Global Switch Cost-1 for RT (Panel D) and composite scores for inhibitory control for the Heart and Flower (or Dots) task (calculated from mean RT; panel E) and composite scores for cognitive flexibility for the Heart and Flower (or Dots) task (calculated from mean RT; Panel F) and composite scores for inhibitory control across first two tasks (calculated from mean RT; panel G) and composite scores for cognitive flexibility across first two tasks (calculated from mean RT; Panel H)

Stress tended to facilitate performance of men and women with low E2 but tended to impair performance of women with high E2, though no significant effects of group or significant pairwise comparisons were found. M=men. F-L=women tested during early follicular phase (low E2 and P4 levels). F-H=women tested during midluteal phase (high E2 and P4 levels). Mean and error bars are based on adjusted statistics. Error bars indicate standard error.
3.4.4 Raven’s Advanced Progressive Matrices:

To control for the effect of differences between versions, total scores were converted into z-scores for each version. Controlling for the effects of stress order (whether the stress or calm session came first) and perceived stress during over the course of the month before testing, the mixed design ANOVA revealed that there was a significant main effect of condition \( (F (1, 65) = 6.79, p = 0.01, \eta^2 = 0.01) \) with significantly worse performance found in the stress condition than in the calm condition with all subjects included in the analysis. However, no significant difference between the more stressful versus calmer conditions was found for any subject group alone (see Figure 18). There was a significant main effect of subject group \( (F (2, 65) = 4.88, p = 0.011, \eta^2 = 0.13) \). Bonferroni-Holm-corrected pairwise comparisons showed that men performed significantly better than women tested when their E2 levels were low \( (t (44) = 3.12, p = 0.008) \).

For the F-L group, performance of well-performing group was significantly more impair than by stress the performance of poorly-performing group \( (t (18) = 2.52, p = 0.021) \). The effect of stress on performance of the F-H or M group did not differ significantly between the well-performing groups and poorly performing groups.

The heart rate after EF3 relative to baseline was negatively correlated with the performance (z scores) for the F-L group \( (r = -0.357, p = 0.024) \) but not for the other two subject groups or for all three groups combined (all \( ps > 0.11 \)). Performance was not
significantly correlated with cortisol level after EF3 relative to baseline for any subject group or for all three groups combined.

Figure 18 Performance (z scores for total scores) on Raven’s Advanced Progressive Matrices by subject group and condition.

Performance was significantly worse in the stress condition than in the calm condition. Across both conditions, our M group scored higher than our F-L group. M = men. F-L = women tested during early follicular phase (low E2 and P4 levels). F-H = women tested during midluteal phase (high E2 and P4 levels). All = all participants. Error bars indicate standard errors. * p<0.05, ** p<0.01.
3.4.5 Re-Ordering Digit Span task (verbal WM):

For longest span, the effect of subject-group, the effect of condition (more stressful versus calmer), and the condition by subject-group interaction all failed to reach significance (all $Fs < 2.2, ps > 0.1$), which might be due to the ceiling effect: There were only four possible span lengths (4, 5, 6, 7; 7 was the upper limit). Although the distribution was symmetric, 20% and 25% of participants reached the upper limit in the stressful and calmer conditions. Unexpectedly, women tested when their E2 levels were high tended to achieve longer spans when stressed than when calm, whereas women tested when their E2 levels were low tended to show shorter spans when stressed than when calm. Men showed an even smaller difference on this measure of WM comparing the more stressful and calmer conditions (see Figure 19A).

3.4.6 Visuo-Spatial WM task:

As with the verbal WM task, the ANOVA for the longest span did not show any significant differences, which might be due to the same ceiling effect as well. The same trends that were evident on the verbal WM task were evident here and in the composite scores across these two WM tasks (see Figure 19B and 19C).
(B)
Figure 19 Mean longest span for the Re-Ordering Digit Span task (A) and our Visuo-Spatial WM task (B) and composite scores across these two tasks (C).

No significant results were found from ANOVAs for either of these WM tasks, but opposite trends are evident between women with high and low E2 levels. M=men. F-L=women tested during early follicular phase (low E2 and P4 levels). F-H=women tested during midluteal phase (high E2 and P4 levels). Error bars indicate standard errors.
4. Discussion

In this study, I investigated the potential influence of sex (specifically the sex hormone, E2) on the effects of mild social evaluative stress on EFs, which are critical for reasoning, problem-solving, self-control and planning. Based on Diamond’s hypothesis that baseline PFC DA levels are higher and closer to optimal in women when their circulating E2 are elevated, we predicted that mild stress, which increases PFC DA levels, should facilitate EF performance in men but impair it in women when their circulating E2 levels were higher. This would extend previous findings from studies of rodents (Shansky et al., 2004; Shansky et al., 2006) and from studies of humans showing that a different way to increase PFC DA levels (Individuals homozygous for COMT-Met\textsuperscript{158} versus Individuals homozygous for COMT-Val\textsuperscript{158}) enhanced performance in men and women when their E2 levels were low but not for women when their E2 levels were high. The results we predicted would also support the role of E2 in one of neurochemical pathways (the PFC DA system), which is involved in EFs.

In line with our predictions, there is some evidence that our social-evaluative stress procedure succeeded in inducing physiological and subjective stress responses, and there is some evidence that it exerted opposing effects on EF performance in men and in women when their E2 levels were elevated.

4.1 The effectiveness of our stress induction procedure

For most physiological indicators of stress (systolic and diastolic BP, HR, and cortisol...
level) and for self-reported stress levels, there were no significant differences at baseline between the more stressful versus calmer sessions. Our aim was to induce mild social-evaluative stress. Unlike studies using TSST paradigm, we did not find a substantial increase in cortisol levels or HR in the stress session\textsuperscript{12}, which may indicate that our stress procedure was not stressful enough. However, for all our physiological indicators of stress except the ratio of cortisol to DHEA levels, either a significant main effect of condition or a significant time by condition interaction, or both, was found in all cases indicating significantly greater stress in the stress condition than in the calm condition. Thus, converging evidence from indicators of HPA axis activity (cortisol) and indicators of SNS activity (HR and systolic and diastolic BP) consistently pointed to our stress induction procedure having been effective in evoking more stress or arousal than participants experienced in our calm condition.

The non-parametric tests of subjective stress also indicated that participants felt significantly more stressed by the end of the stress session than after the calm session. It is worth noting that a study using DA D2/3 PET found that greater subjective stress from a psychosocial stress-induction was accompanied by greater DA release in PFC (Lataster et al., 2011), which is consistent with Cerqueira et al. (2007b) and Roth et al. (1988).

\textsuperscript{12} Indeed, one committee member suggested that our social-evaluative stress procedure was perhaps more arousing than stressful.
4.2 Sex differences in stress responsivity

Men in our study had significantly higher baseline systolic BP than either group of women. There was no significant condition X subject-group or condition X time X subject-group interaction, however, indicating the effects of our stress induction procedure on responses of systolic BP between did not differ among the three subject groups.

Unlike some studies utilizing the TSST (Kirschbaum et al., 1999; Schoofs et al., 2013b), we did not find stronger cortisol responses in men than in women. Instead, although all three subject groups suggested a greater difference in later cortisol levels compared to baseline in the more stressful than in calmer condition throughout testing, at the after EF3 timepoint, women with low E2 levels showed a relatively larger cortisol increase from baseline than women with high E2 levels. No other sex difference in stress reactivity measures was found than mentioned above. In addition, there were few significant correlations between HR or cortisol levels relative to baseline levels and EF performance for any subject group. Thus, the observed sex differences in EF performance probably cannot be attributed to sex differences in the degree of stress or arousal elicited by our social-evaluative stress procedure.

4.3 Sex differences in the effects of stress on executive functions

Although opposing effects of stress in females and males on cognitive functions dependent on PFC (Shors et al., 1998), including PFC-dependent EFs (Shansky et al., 2004; Shansky et al., 2006), have long been known in rodents, only a few studies have investigated this in humans. Two studies, using the TSST as the acute stressor, both found significant
subject-group X condition interactions in performance on the n-back task, with men showing
stress-induced enhancement and women showing no difference between the more stressful
and calmer conditions or a trend towards stress-induced impairment (Cornelisse et al., 2011;
Schoofs et al., 2013). However, as mentioned previously, they did not control for COMT
genotype or E2 levels (or menstrual phase). Moreover, the majority of female participants in
one study used oral contraceptives, which are known to not only affect sex hormones but also
dampen stress responsivity (Cornelisse et al., 2011). These might be the reasons why they
failed to find significant stress effects on women’s performance.

**Inhibitory Control.** In the present study, the finding that inhibitory control was
significantly impaired when stressed (indicated by a larger difference in % correct between
incongruent trials and easy [congruent or no distractors] trials in the stress condition than in
the calm condition) in women tested when their E2 levels were high but not in men or in
women tested when their E2 levels were low is consistent with animal studies on the
modulating role of E2 in stress-induced EF impairments in females (Shansky et al., 2004;
Shansky et al., 2006). One recent study with human participants that found that acute stress
(induced by a social-evaluative cold pressor test [SECPT]) enhanced inhibitory control
performance on the stop-signal task in both men and women; they found no sex difference in
the effects of stress on performance(Schwabe et al., 2013). The reasons why their findings on
stress effects in women differ from ours might be: 1) They did not control for E2 levels or
menstrual phases in women. They only excluded women using oral contraceptives or women
who were in their menses. 2) Different paradigms which tapped different aspects in
inhibitory control were used. EFs, even each core EF, represent a set of mental functions and
can not be viewed as a unitary one (Floresco & Jentsch, 2011). Each aspect of EFs may be subserved by anatomically distinct brain circuits and follow different DA dose-response curves. Both of our first two tasks required inhibiting one response in order to make another, whereas the stop-signal task requires inhibiting a response one was about to make in order not to respond at all (Diamond, 2013). A recent meta-analysis indicated that the direction of stress effects on inhibition depends on the type of inhibition (Shields et al., 2016).

Specifically, stress significantly facilitated response inhibition but impaired interference inhibition. 3) The lengths of intervals between the stress induction and the EF task were different: Because they were interested in the effects of cortisol, the stop-signal task was delivered 30 mins after the stress induction (the SECPT), whereas our stress induction occurred during our EF tasks, and our first two EF tasks were administrated within 20 mins after the onset of stress. Neurochemical systems triggered by acute stress have different temporal signatures (the release of catecholamines is rapid, the release of cortisol is slow; Hermans et al., 2014). Around 30 mins after the onset of stress, cortisol levels may reach their peak but the stress-induced release of catechoamines might have decreased towards basal levels by them. Different levels of these neuromodulators at different intervals after the onset of stress may lead to different changes in EF performance. 4) Unlike our study, they used between-subject comparisons to detect stress effects where individual differences might affect results; we used within-subject comparisons.

**Cognitive Flexibility.** Cognitive flexibility, the core EF which builds on inhibitory control and working memory, includes the ability to switch between different ways of thinking or different tasks or rules. The mixed blocks in our first two tasks assessed this core EF by
requiring participants to flexibly switch between two rules on a trial by trial basis. We found a trend towards stress-induced improvements in men but a trend towards stress-induced impairments in women tested when their E2 levels were high on the indices measuring local switch cost (RT difference between switching trials and non-switching trials in the mixed block) and global switch cost (RT difference between all trials [or incongruent trials] in the mixed block and all trials [or incongruent trials] in the single-task blocks). However, the group differences failed to reach significance.

Interestingly, compared to the calm condition, participants in the stress condition showed significantly shorter RTs on switching trials in the mixed block of our Flanker/Reverse Flanker task. Such shorter RTs in the stress condition could not be attributed to a speed-accuracy trade-off because there was no significant main effect of stress in analyses of accuracy.

These findings on cognitive flexibility are inconsistent with some previous human studies (Alexander et al., 2007; Plessow et al., 2011). In a study by Alexander et al. (2007), the Compound Remote Associates test (CRA) and the anagram tasks (both of which require cognitive flexibility) were administrated after the TSST or in a control condition. Both men and women showed impaired performance with fewer correct responses on the CRA test and longer latency for the anagram test. The explanations for the discrepancies between their results from ours might include: 1) Our study and theirs use different paradigms that required different aspects of cognitive flexibility—our tasks require flexibly switching rules whereas the CRA test and the anagram task require solving problems in new ways, which may show different dose-response curves; 2) Our study and theirs use different stress induction
procedures. A recent meta-analysis revealed that as stress severity increases, a greater stress-induced impairment on cognitive flexibility is found (Shields et al., 2016). Our stressor was milder than the TSST they used, as reflected by physiological measures. Unlike ours, their stress procedure induced a substantial increase of HR during the stress condition, which might be associated with supraoptimal levels of arousal and of stress-induced catecholamines. 3) As mentioned previously, like most human studies on stress effects, they did not control for COMT genotype, which affects basal PFC DA levels and influences the direction of the stress effects.

**Working Memory.** For WM, although both our first two tasks require it, they do not permit teasing it apart from inhibitory control or switching, so we used two other tasks (Re-Ordering Digits and Spatial Span) to assess working memory (but they were introduced after ~40% participants had already been tested). Although the directions of stress effects were the same for both of these two tasks for each subject group, no significant stress effects were found in within-group comparisons across the two conditions. Contrary to our predictions, a trend towards stress-induced improvements was found in women tested when their E2 levels were high (and to a less extent also in men), whereas women tested when their E2 levels were low showed a trend towards stress-induced impairments. It is important to note that, although the stressor always occurred during the two WM tests, these tests occurred 60 - 85 mins after the onset of stress. Participants’ stress levels and neuromodulator levels might have differed here from those present during the first two tasks, which might explain differences in the direction of stress effects on these tasks. That said, with a larger sample size and a higher upper limit of span length, the stress effect on WM performance might be
greater and more reliable.

**Fluid Intelligence/ Reasoning.** Fluid Intelligence, one of the higher-order EFs, was assessed using RAPM, which requires reasoning. Contrary to our predictions and our findings on the first two tasks, stress impaired the performance of all three groups. There are three potential explanations for this: 1) Higher-order EFs (i.e. fluid intelligence) might recruit a different and larger neural network that may have a different dose-response curve for PFC DA levels. The hippocampus has been shown to be essential for successful performance of RAPM (Zhu et al., 2014). Stress and GC increase have been repeatedly shown to hamper hippocampus-dependent cognitive functions (Lupien et al., 2007). 2) In our study, the RAPM was administered around 30 - 50 mins after the onset of stress. As mentioned earlier, levels of neuromodulators during RAPM might be different from those when participants were doing our first two tasks. One of the primary differences might have been in levels of cortisol. Cortisol levels are known to peak around 30 mins after the onset of stress. Our data also show that the stress-induced increase of cortisol was higher after RAPM than after the first two tasks. The impairing effects of elevated cortisol on EFs have been demonstrated in many human studies (e.g. Elzinga & Roelofs, 2005). 3) Although our stressor always occurred during testing, the stressor might haven been more noticeable during the 25-min long exam-like paper-based RAPM than the first two computerized tasks with colored stimuli and cheering sounds at the end. As a result, participants might have experienced more stress or greater cortisol responses to stress during the RAPM than during the first two tasks, which might have led to above-optimal levels of PFC DA.
4.4 Potential underlying mechanisms

The inverted U-shaped dose-dependent relationship between levels of PFC DA and EF performance via DA D₁ receptor (D1R)-cyclic adenosine monophosphate (cAMP) signaling has been established in animal studies (see (Arnsten, Wang, et al., 2015) for a detailed review). Thus, baseline DA levels are essential in determining the direction of stress effects on EF performance. Diamond (2007) hypothesized that, due to E2 reducing COMT enzymatic activity in PFC, baseline levels of PFC DA levels would be higher and closer to the optimal in women when their E2 levels are elevated than in men. As it has been repeatedly demonstrated in rodents, monkeys and humans that even mild stress increases PFC DA levels (Cerqueira et al., 2007b; Vaessen et al., 2015), it seemed reasonable to predict that the effect of stress-induced elevation of PFC DA would differentially affect men and women during menstrual phases when their circulating E2 are elevated. Our findings show that mild stress tend to impair inhibitory control and cognitive flexibility in women tested during the midluteal phase (high E2 and P4 levels) whereas inhibitory control and cognitive flexibility of men and women tested during the early follicular phase (low E2 and P4 levels) showed a trend toward stress-induced improvements, which supports the model shown in Figure 1.

However, we do not claim that this is a complete picture of mechanisms underlying the observed sex differences in the effects of stress on EFs. Other prime candidates include: 1) E2 may modulate other neurochemical pathways that are involved in stress effects on EFs. First, E2 may interact with the locus coeruleus (LC)-NE system. Exposure to stress activates
the LC-NE system and increases central NE release and specifically NE in PFC, possibly via stimulation by corticotropin-releasing factor projections from amygdala (Arnsten, 2009a; Van Bockstaele et al., 1998), which is followed by a global increase of NE via the SNS activation. Animal studies have intensively studied the role of the LC-NE system in stress effects on EFs (Arnsten, Raskind, et al., 2015; Berridge & Spencer, 2016). In a human study, pharmacological modulation of the LC-NE system has been shown to reverse the stress-induced impairments on cognitive flexibility, indicating a modulating role of the LC-NE system on stress effects on EFs in humans (Alexander et al., 2007).

The interaction of E2 with the LC-NE system in stress effects on EFs was revealed by a rodent study showing that stimulation of NE alpha-2 adrenoceptors reversed stress-induced WM impairments in male rats and OVX rats but not in OVX rats with E2 replacement (Shansky et al., 2009).

Second, although not supported by our data, some rodent and human studies have demonstrated that E2 can modulate HPA stress responsivity and stress-induced GC release (Kudielka & Kirschbaum, 2005; Mitsushima et al., 2003). GCs (mainly cortisol in humans) have been shown to play a critical role in stress effects on EFs. Schwabe et al. (2013) has demonstrated that blocking the mineralocorticoid receptor, one type of receptors for cortisol, blocked the stress-induced enhancement in response inhibition in both men and women. A rodent study showed that if you block GRs in PFC, you do not see stress-related DA release in PFC of male rats and similarly, if you administer GC directly to PFC locally, you see a greater stress-induced DA release in PFC in male rats (Butts et al., 2011). Moreover, it has
also been suggested that GCs can block extraneuronal catecholamine transporters on central nervous system glia to further increase extracellular catecholamine levels (Gründemann et al., 1998). Therefore, an E2-mediated larger increase of cortisol levels might push PFC DA levels further past the optimal point on the inverted U-shaped curve, which would lead to stronger, more negative stress effects in women when their E2 levels are high. 2) Men and women may use different strategies, calling upon different brain circuits, to perform a given EF task. An imaging study on sex difference in solving Raven’s Matrices provided evidence that the cortical regions associated with comparable scores on Raven’s Matrices differed between men and women (Yang et al., 2014). In this case, the effects of stress on different brain regions might lead to sex-dependent stress-induced changes. Future studies with the help of imaging techniques could determine whether different brain circuits are involved when men and women perform the tasks where we found sex differences and could investigate how E2 plays a role in that.

4.5 Limitations

The large inter-individual differences in the length of menstrual cycle is mostly due to the variation in the length of follicular phase, whereas the luteal phase is relatively stable with 14 days from ovulation to the onset of menses (Hampson & Young, 2008). To better estimate the phase of the menstrual cycle characterized by high E2 levels, we chose the midluteal phase (moderately high E2 levels and high P4 levels) instead of late follicular phase (high E2 levels), which brings up consideration of the influence of P4. It is important
to note that, like data from our participants, others have found that midluteal P4 concentrations were approximately 100-fold greater than midluteal E2 levels (Sundström Poromaa et al., 2014). A post-mortem study indicated the presence of P4 in different brain including the frontal cortex (Bixo et al., 1997). Although it has not been demonstrated in neurons yet, a cell culture study revealed that P4 can either up- or down-regulate COMT gene expression depending on which P4 receptor isoform is present (Salama et al., 2007). In addition, intracerebroventricular injection of allopregnanolone, one of the metabolites of P4, has been demonstrated to decrease both basal and stress-induced PFC DA release in rodents (Motzo et al., 1996). Taken together, although the study on OVX female rats showed only E2 replacement but not P4 replacement caused the comparable stress-induced impairments to those seen in female rats in proestrus (Shansky et al., 2004), the potential roles of P4 and its metabolites in modulating the effects of stress need to be further investigated.

In addition, as mentioned earlier, there are also large inter-individual differences in E2 and P4 in women during each menstrual phase (Sundström Poromaa et al., 2014). Some women might have higher E2 levels at the early follicular phase than other women have during their midluteal phase. To better check whether female participants were tested during the menstrual phases they reported, it would be wise to compare within-subject hormone levels during both the self-reported early follicular phase and midluteal phase. Future studies should attempt to better monitor the hormonal levels of female participants or consider a longitudinal design to establish the role of E2.

The single nucleotide polymorphism (SNP) of COMT genotype, Val^{158}Met, was
controlled for in this study, however, there are other genetic factors that affect EFs in humans which should perhaps be controlled for in future studies. The 2-SNP haplotype of the COMT gene has been demonstrated to have stronger effects on WM than those observed for single SNPs (Meyer-Lindenberg et al., 2006). One recent study has demonstrated that SLC6A4 genotype, which encodes the serotonin transporter, interacts with stress from one’s mother’s mood, determining whether such stress did or did not impair the EF performance of six year-old children (Weikum et al., 2013).

Moreover, as mentioned earlier, our social-evaluative stress procedure did not elicit substantial cortisol or HR response, indicating the procedure might not have been stressful enough. Future studies with improved stress procedure might yield more significant results. Another concern in the present study is the effectiveness of our stress induction procedure in inducing the release of PFC DA at the individual level. It has been frequently reported that there are large inter-individual variations in stress responses to laboratory psychological stressors (Dickerson & Kemeny, 2004; Lupien et al., 2007). Depending on one’s sex, motivation, coping strategies, goals, genetic endowment, level of stress in one’s life, and emotional state, stress could be appraised differently and elicit different affective states, neurochemical responses and behavior (Dickerson & Kemeny, 2004; Salvador, 2005; Vaessen et al., 2015). For research investigating effects of stress-induced cortisol levels on EFs, a study revealed that cortisol-responders showed stress-induced impairments on WM performance whereas cortisol-non-responders showed stress-induced improvements (Elzinga & Roelofs, 2005). In the current study, we were unable to check directly whether our stress procedure truly induced elevations of PFC DA release. For stress-induced PFC DA release,
researchers usually use PET or single-photon emission computed tomography with DA D1 and D2/3 radioligands to measure DA release in living human brain (Guo et al., 2003; Laruelle, 2000). Based on results from these imaging studies, the stress indicators used in our study do not reliably correlate with stress-induced PFC DA release. Different imaging studies using the same stress induction procedure, the Montreal imaging stress task (a modified version of the TSST, have found significant associations between different stress indicators and stress-induced increases in PFC DA levels. Mizrahi et al. (2013) found a significant positive association between stress-induced PFC DA release and cortisol levels, but two other studies found no association between those (Lataster et al., 2011; Nagano-Saito et al., 2013). Instead, one of them found stress-induced increase of PFC DA levels was significantly associated with subjective stress levels (Lataster et al., 2011), but the other found the association to be with HR (Nagano-Saito et al., 2013). In addition, human imaging studies have also shown that the magnitude of stress-induced release of DA seems to be dependent on early life experiences, personality traits and genetic variations in several SNPs (Vaessen et al., 2015). Thus, whether our stress procedure truly induced elevations of PFC DA release in our participants is unknown. To ensure the effectiveness of stress induction procedures and to get more reliable information on underlying mechanisms, future studies with the help of imaging techniques are warranted.

4.6 Implications

As far as we know, the present study is the first to investigate potentially gonadal
hormone-mediated sex differences in the effects of mild acute stress on EFs in humans. We tried to control somewhat for baseline PFC DA levels by including only COMT gene heterozygotes. Our findings indicate that mild acute stress tended to impair inhibitory control and cognitive flexibility (task switching) in women tested when their circulating E2 levels were high (the midluteal phase) but not in men or in women tested during their early follicular phase when E2 levels are low. It provides evidence consistent with a modulating role of E2 on the PFC DA system, affecting an underlying mechanism by which stress affects cognition. The findings also indicate that regardless of sex or gonadal hormone levels, mild acute stress seems to be harmful for reasoning. One might conclude from our results that to achieve a better life, it might be worthwhile to learn and practice stress management strategies, such as meditation, yoga, or physical exercise, and this might be especially beneficial for women when their E2 levels are elevated. It suggests that while men may need some stress to perform at their best, women can often perform quite well without it. Different people thrive under different conditions; those who look awful in one environment might shine in another. For example, might we be missing giftedness in some women by assuming that keeping people a little on edge is beneficial for all? A man might be procrastinating or taking risky chances to try to correct below optimal PFC DA levels without realizing it; once the cause is correctly identified more productive fixes might often be found.

Moreover, for drugs which affect PFC DA levels, the patient’s sex, genotypes, and hormonal information may need to be taken into account to prescribe the proper doses. In addition, with significantly higher incidence rates of stress-related mental illnesses being reported in women than in men, such as depression and anxiety disorders (Cover et al., 2014),
these results provide information for the potential roles of hormones in development, as well as preventive and therapeutic implications for these illnesses.
5. Conclusion

In this study, consistent with our predictions, we found that mild acute stress differentially affected EF performance in men and women depending on a woman’s menstrual phase (i.e. E2 and P4 levels). Women tested when their E2 levels were high tended to show stress-induced impairments, whereas a trend towards stress-induced improvements (or no effect of stress) was found in men and in women tested when their E2 levels were low. Our findings extend those from animal studies and COMT gene studies in humans and support the modulating role of E2 on the PFC DA system. However, our findings also differ from some previous human studies on the effect of stress on EFs which did not find a sex difference.

Such differences might be due other studies (a) not taking into account COMT genotype or E2 or P4 levels, (b) using different EF paradigms and different stress paradigms, and (c) imposing a different interval between onset of stress and EF tasks.

Contrary to our predictions, stress-induced impairments were found for both men and women on our one measure (Raven’s Progressive Matrices) of higher-order EFs, fluid intelligence/reasoning.

Future studies with the help of imaging techniques should attempt to further elucidate the neurochemical mechanisms underlying stress effects on EFs. In addition, our findings suggest that ways of managing stress, improving EFs and achieving a better life need to be considered in a sex-specific and hormone-dependent manner.
Bibliography


memory: associations with differing cognitive operations. *Biological Psychiatry*, 58(11), 901-907. doi:10.1016/j.biopsych.2005.05.010


doi:10.1016/j.neuropsychologia.2011.11.003


doi:10.1016/j.neuropsychologia.2010.04.023


Gozansky, W. S., Lynn, J. S., Laudenslager, M. L., & Kohrt, W. M. (2005). Salivary cortisol determined by enzyme immunoassay is preferable to serum total
cortisol for assessment of dynamic hypothalamic--pituitary--adrenal axis activity. 


Guo, N., Hwang, D. R., Lo, E. S., Huang, Y. Y., Laruelle, M., & Abi-Dargham, A. (2003). Dopamine depletion and in vivo binding of PET D1 receptor radioligands: implications for imaging studies in schizophrenia. *Neuropsychopharmacology, 28*, 1703-1711. doi:10.1038/sj.npp.1300224


Appendix

Subjective Assessment of Stress Immediately Before (A) and After Testing (B)

Subject ID# 
Testing session#

Stress Assessment

Please circle the number corresponding to how you feel at this moment

**Right now, I feel..........................**

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>Very relaxed</td>
<td>Pretty relaxed</td>
<td>Slightly Relaxed</td>
<td>Slightly Stressed</td>
<td>Pretty Stressed</td>
<td>Very Stressed</td>
</tr>
</tbody>
</table>

Additional comments concerning any stressful event that has occurred today:___________________________

_____________________________________________________________________________________

_____________________________________________________________________________________

_____________________________________________________________________________________

(A)
Post-Activity Assessment

Please circle the number corresponding to how you felt

During the cognitive testing I felt.........................

1 2 3 4 5 6
Very relaxed Pretty relaxed Slightly Relaxed Slightly Stressed Pretty Stressed Very Stressed

Additional comments concerning any stress experience during the testing__________________________

________________________________________________________________________________________

________________________________________________________________________________________

________________________________________________________________________________________

(B)