From The Department of Clinical Neuroscience
Karolinska Institutet, Stockholm, Sweden

INVESTIGATING AND EXPLAINING SEX DIFFERENCES IN EPISODIC MEMORY

Martin Asperholm

Karolinska
Institutet

Stockholm 2020

All previously published papers and figures were reproduced under open licenses or with permissions from the respective publishers.

Published by Karolinska Institutet.

Printed by Universitetsservice US-AB

ISBN 978-91-7831-771-4

INVESTIGATING AND EXPLAINING SEX DIFFERENCES IN EPISODIC MEMORY

THESIS FOR DOCTORAL DEGREE (Ph.D.)

By

Martin Asperholm

Principal Supervisor:
Professor Agneta Herlitz
Karolinska Institutet
Department of Clinical Neuroscience
Division of Psychology

Co-supervisors:
Professor Håkan Fischer
Stockholm University
Department of Psychology

Professor Gustav Gredebäck
Uppsala University
Department of Psychology

Opponent:
Professor Gijsbert Stoet
University of Essex
Department of Psychology

Examination Board:
Dr. Rickard Carlsson
Linnæus University
Department of Psychology

Professor Timo Mäntylä
Stockholm University
Department of Psychology

Professor Guy Madison
Umeå University
Department of Psychology

ABSTRACT

Sex differences can be seen in a lot of different areas, one of them being cognition where, among other things, men tend to perform better at more spatial tasks and women at more verbal. When it comes to episodic memory — the memory for past events in terms of the what, where, and when — there has been some circumstantial evidence of sex differences. In Study I of this thesis, we performed a meta-analysis on mean sex differences in episodic memory, using a dataset of 617 studies, totaling over one million subjects. This dataset consisted both of published and unpublished data, as well as several open databases. Here, the main finding was that, just like with other cognitive tasks, males tend to perform better on more spatial tasks while females tend to perform better on more verbal tasks. It could also be shown that females performed better when remembering faces, as well as tasks having to do with smell, touch, and different shades of colors.

Further, it has been shown that males tend to be more variable than females for a lot of different traits, both biological and psychological. Even if men, on a group level, usually have larger variance when it comes to cognition in general, there is less support for saying anything about episodic memory. In Study II of this thesis, we performed a number of analyses on a somewhat reduced version of the dataset gathered in Study I, searching for variance differences between males and females. Here, there were some evidence showing that males were slightly more variable than females. However, through exploratory investigations we also found results suggesting that this pattern potentially could come about because of underlying methodological problems in the original research.

Finally, there have been some large-scale studies suggesting that when it comes to improvements in cognition, women tend to benefit more than men from social progress and increased living conditions in a society. There has also been some evidence for this relationship when it comes to episodic memory. In Study III, we expanded upon these findings by examining the dataset gathered in Study I, which comprises more countries and a larger timeframe than any other investigation on this topic. Here, we could show that for verbal episodic memory, sex differences were related to several different indicators of living conditions tied to the year and country of each study. However, when pitted against each other, it was only the overall education and employment

level that could be shown to be related to the outcome and not gender equality, which we expected would be the most important indicator.

In this thesis, I first present a comprehensive background on the topics presented above, including some in-depth, possible explanations for why things are the way they are from evolutionary, biological, and social perspectives. I then go through the dataset that all three studies are based on, as well as present the result from them, including some of my own interpretations. Finally, I discuss some general themes that relate to the studies performed, including necessary choices when collecting data for meta-analyses, possible bias in the dataset, and statistical considerations when dealing with dependent data.

LIST OF SCIENTIFIC PAPERS

I. Asperholm, M., Högman, N., Rafi, J., & Herlitz, A. (2019). What did you do yesterday? Sex differences in episodic memory. *Psychological Bulletin*, *145*(8), 785–821. doi:10.1037/bul0000197

II. Asperholm, M., van Leuven, L., & Herlitz, A. (2020). Sex differences in episodic memory variance. *Frontiers in Psychology*, *11*(613), 1–10. doi:10.3389/fpsyg.2020.00613

III. Asperholm, M., Nagar, S., Dekhtyar, S., & Herlitz, A. (2019). The magnitude of sex differences in verbal episodic memory increases with social progress: Data from 54 countries across 40 years. *PLOS ONE*, *14*(4), 1–11. doi:10.1371/journal.pone.0214945

CONTENTS

1 INTRODUCTION

1.1 OVERVIEW OF THE DISSERTATION

For my thesis, I am going to investigate the general patterns, moderators, and potential explanations of sex differences in episodic memory performance from a cognitive and social perspective. I will examine this theme through three separate studies, all of them applying the meta-analytic method in some form on the same, very large, dataset of articles, unpublished data, and open databases that carries information on the matter at hand. More specifically, in Study I (Asperholm, Högman, et al., 2019), we[1] investigated the overall sex differences in mean episodic memory performance. Next, in Study II (Asperholm et al., 2020), we turned to examine the overall sex differences in episodic memory variance. Finally, in Study III (Asperholm, Nagar, et al., 2019), we investigated how mean sex differences in episodic memory might vary together with factors pertaining to social progress and living conditions in society.

1.2 DEFINITION OF SEX

When I talk about *sex* in this thesis, I refer to the biological/anatomical division of human beings into men and women. This can be contrasted with *gender*, which normally is used to refer to more mailable, social and/or psychological aspects, for example gender identity and gender roles. So the underlying concept that I am trying to capture here is a binary division of people into males and females from a biological/anatomical point of view for all cases where it is obvious which category someone belongs to (which is true for the overwhelming majority of them). Edge cases when it is unclear which sex a certain person belongs to, is not included in my definition of sex and will not be addressed.[2]

[1] Throughout this thesis, I will often use the terms “we”/“our”, not as the author/editorial/royal *we* (even if that also will happen) that often is used in, for example, computer science, but rather as a way of referring to processes, decisions, and conclusions that were carried out in the context of our studies. Contrary, when speaking of more personal thoughts and reflections (for which my co-authors shall not be held responsible), I will instead try to use the concept I just used: “I”.

[2] This was rather the subject of one of my research group colleges’ Ph.D. thesis (Strandqvist, 2018).

It should be noted that in most psychological studies, including those that form the basis of this thesis, sex is registered by either letting the participants fill it out themselves or by having the experimenter make this judgment. This makes room for both mistakes as well as answers based on other definitions of sex. However, I am going to assume that the categorization performed in this manner comes sufficiently close to if biological/anatomical examinations *would* have been carried out for all of the participants. Results from these studies should therefore be able to inform research questions pertaining to the definition of sex that I am interested in.

1.3 WHY STUDY SEX DIFFERENCES?

Before embarking on the research at hand, we should first take a step back and reflect on why sex differences — and more specifically, sex differences in cognition — should be investigated in the first place. What is clear is that all types of research can be pitted against all other types of research in trying to discern what the best use of non-infinite resources are: Is understanding the mechanisms behind aging more important than elucidating the processes contributing to climate change? Should one focus on AI safety[3] rather than building computational models of vision? Would curing cancer be more advantageous than figuring out how to stop viral proliferation?[4] Answers to questions like these will, of course, be dependent on personal values as well as subjective beliefs about what the research in question might (or might not) lead to, so trying to rank the topic at hand against others would be of limited value here. However, it *is* possible to list the merits of a certain research field independent of its relative value.

I can personally identify three major reasons for why it is important to study sex differences:

> **Estimating baselines.** If one does not know *that* and *to what extent* there are sex differences for a certain trait, one cannot estimate reasonable baselines for manifestations of that trait either. For example, if one did not know that and to what extent women

[3] If you wonder what AI safety even *is*, Nick Bostrom's book *Superintelligence: Paths, Dangers, Strategies* from 2014 is a good primer.

[4] A question which unfortunately is quite topical at the time of me writing this with the SARS-CoV-2 virus just having caused the COVID-19 disease to go pandemic (World Health Organization, 2020).

generally score better than men on tests of verbal performance (something that we will get to in section 1.6.2), one might also draw the conclusion that boys somehow are discriminated against after observing that girls tend to get higher school grades in language subjects, even though this outcome might be exactly what is expected.[5]

Deriving underlying reasons. If one *does* know that and to what extent there are sex differences for a certain trait but not *why*, one might, again, draw faulty conclusions as to why the difference exists. For example, one might be aware of the general tendency for men to be more promiscuous than women. However, if one is not aware of the evolutionary theory stating that women have more to lose from an unwanted pregnancy than a man, one might also draw a different conclusion for why the difference exists, for example because of social norms. The upshot of this and the previous concept is that if one does not know that, to what extent, and why a certain sex difference is present, one might also put resources into trying to combat it that might be misdirected or even counterproductive. For example, Stoet and Geary (2013) propose that if policymakers wanted to minimize the sex differences between boys and girls when it comes to reading and mathematics (an aspect that will be further explored later in this introduction), based on the data, they should focus on helping the lowest-performing boys in reading and highest achieving girls in mathematics. Reaching this conclusion requires knowledge about several aspects of what the sex differences actually look like.

Easily implementable actions. Research on sex differences has the potential to yield easily evaluated and implementable actions in areas such as mental health care and community planning. For example, if it can be shown (which it, by the way, can; Rosenfield and Mouzon, 2013) that men and women tend to differ when it comes to mental health problems or how they respond to a certain psychological treatment, this could also help design proper screening instruments and preventive measures. Similarly, knowing that boys and girls tend to differ with regard to certain cog-

[5] As a technical side note, it is important to know more than just simply mean differences between two groups when trying to estimate different outcomes. This will be further discussed in section 1.5.

> nitive traits might help to design the school curriculum in a more beneficiary way. While the same thing can be said for many other types of categorizations as well, research on sex differences has the benefit that the division often tends to be relevant in many different fields and also that it is relatively simple to carry out. That is, while many groupings require some kind of extra data collection (for example, categorizing people into introverts and extroverts requires them to take a test), sex is easy to determine and is even often already registered anyway.

However, one should also reflect on possible downsides. That certain type of research is being conducted can sometimes be problematic or controversial, something that in recent times was highlighted when a Dutch research group wanted to, and subsequently also did, publish findings of how they managed to create an airborne version of A/H5N1 (more commonly called the *avian influenza*) among European polecats (Herfst et al., 2012; Russell et al., 2012). This publication decision was something that was preceded and subsequently followed by a lively discussion, and also actual rule changes in many countries (Fouchier et al., 2013), mostly surrounding concerns that the knowledge could be used by nefarious organizations to create biological weapons.

Psychological research does not have the capacity for the same type of direct hazardous consequences, unless for participants partaking in ethically questionable experiments like, for example, the Milgram experiment (1963). However, it is still an area that has created a lot of controversies throughout the years, one reason being that it is perceived as something that could change the public opinion on different matters for the worse. An example of this is when racial differences have been studied. This topic has, also in modern times, spurred a lot of heat merely when people have presented theories about why things are the way they are, one example of this being the aftermath (Harris, 2017) of the publishing of *The Bell Curve* (Herrnstein & Murray, 1996), a book in which the authors concluded that racial differences in IQ probably are dependent on both environment and genetics, although they remained agnostic to what extent each factor might contribute.[6]

[6] Here is one quote regarding the authors' agnosticism, found on page 311 in the book (Herrnstein & Murray, 1996): "If the reader is now convinced that either the genetic or environmental explanation has won out to the exclusion of the other, we have not done a sufficiently good job of presenting one side or the other. It seems highly likely to us that both genes and the environment have something to do with

While a lot of the indignation with respect to *The Bell Curve* has been about the author's possibly ulterior motives rather than being about bad science in and of itself (I have never heard of an astrophysicist being ostracized from the scientific community and public discourse because of poor methodology), my point here is that merely suggesting — even if it just is based on nothing but theoretical arguments about evolution — that race differences in IQ to some extent *could* depend on genetics probably would be met with not only skepticism but also with some amount of hostility and accusations of a hidden agenda. This shows, justifiably or not, how controversial this topic is.

Research on sex differences is in a similar vein something that also has been controversial (Eagly, 1995). One concrete example of this surrounds an article by Hill (2017) outlining a mathematical model meant as a starting point to discuss the theory that males show greater variability than females, a theroy that is further explored in Study II of this thesis. The paper was first rejected after it already had been accepted in one journal, and then deleted (not retracted) after it already had been published in another (Azvolinsky, 2018; Hill, 2018); both decisions apparently driven by concerns of how the paper would be received from a political, rather than scientific, point of view. In the words of Hill (2018) himself:

> In my 40 years of publishing research papers I had never heard of the rejection of an already-accepted paper. And so I emailed [the editor-in-chief of the journal]. She replied that she had received no criticisms on scientific grounds and that her decision to rescind was entirely about the reaction she feared our paper would elicit.

As an interesting parallel, for Study II of this thesis, we had to remove the description of a specific evolutionary explanation in order to get the paper accepted because it was judged by the journal to potentially drive a sexual stereotype (for a reference to the theory having to be removed, see footnote 14 on page 18).

Taking a step back, I can personally identify two major themes regarding possible concerns with research on sex differences:

Prejudices and discrimination. Researching and subsequently

racial differences. What might the mix be? We are resolutely agnostic on that issue; as far as we can determine, the evidence does not justify an estimate."

finding sex differences in different areas can spawn unwarranted prejudices and discriminatory behavior when the results are interpreted erroneously, which often is the case. For example, even if it could be shown without any reasonable doubt that there are large sex differences for a certain cognitive trait, this would hardly make it rational to assume someone's ability for that specific trait from their sex alone. Rather, *even* for effects that are considered very large, individual cases that runs contrary to the finding are extremely common.[7] As a concrete example, a Cohen's *d* (a measure which we will return to in section 2.2.2) effect size of 0.8, which generally is counted as a large one (Cohen, 1988, pp. 24-27), means that when selecting a random person from the excelling sex, there is a 71.4% probability that he or she will outperform a randomly selected member of the other sex (where there would be a 50% probability if there were no sex difference). Even so, many people often tend to interpret found effects way too strong, believing that no further information really is needed about a certain person in order to defer his or her relative ability. Of course, on the flip side, results showing that there are no or much smaller differences where people previously thought there were large ones could help combat the same type of prejudices and discriminatory behavior. My personal opinion on this matter is that one should first and foremost focus on education and scientific communication, exhausting these options to the fullest, before even starting to think about restricting what can and cannot be said.

Misleading division. By conducting research on sex differences, this could send out a signal that the division between men and women is the end-all categorization to understand the causal reasons behind different type of patterns, problems, and behaviors. Here, it could be argued that most differences probably do not depend only on the sex, but rather on some other underlying factors that tend to cluster unevenly in the different sexes. As an example of this, economic negotiation outcomes tend to favor men (Mazei et al., 2015), but it could be argued that the outcome depends on

[7] And realistically speaking, when it comes to sex differences, effect sizes are often not close to being that large. For an illustration of a number of more typical effect sizes, see Figure 9 which depict assumed sex distributions from some of the investigations in this thesis.

> personality traits that members of both sexes can hold but that appear more often in men. In this case then, maybe a better division could be made using some type of personality trait, meaning that the partition into men and women is misleading when it comes to actually understanding what is going on or to better predict it. With that said, it can often be a good first step to investigate and make clear what is happening at the sex level before digging deeper, trying to come up with even better predictors.

In the end, one will have to make a personal judgment on how to weigh these potential benefits and drawbacks associated with research into sex differences.

1.4 MEMORY AND EPISODIC MEMORY

When it comes to memory, there are several different categorical divisions that frequently are being used to delineate various memory processes (see Figure 1 for an illustration). First, one can make a distinction between *sensory memory*, *short-term memory*, and *long-term memory*. Sensory memory is the very short-lived memory, about a second or so, that remains after a sensory experience (Atkinson & Shiffrin, 1968; Sperling, 1960). In order to process this information, content from the sensory memory can be brought into the short-term memory, or *working memory* as it is more popularly called[8], a type of memory where data can be stored and manipulated as long as it is actively thought about (Baddeley, 2018). However, in order to keep this material for future use, it has to be transferred into the long-term memory, a form of memory where the material can remain indefinitely.

Further, within long-term memory, a frequently used division is that between *implicit* and *explicit* memory (Reder, 1996). Implicit memory here refers to the type of memory processes that are non-dependent on the person being cognizant[9] of the process. For example, while you

[8] This distinction has mainly been used to separate the mere process of just storing information (short-term memory), from the process of storing *and* manipulating information (working memory). Simply remembering five numbers would be an example of the former. Remembering these five numbers while also summing them would be an example of the latter.

[9] I am aware that the more conventional word to use here is *conscious*. However, my whole master thesis (Larsson, 2011) was basically about arguing that this word is used in a confusing and ambiguous way, meaning that I cannot, in good faith, use

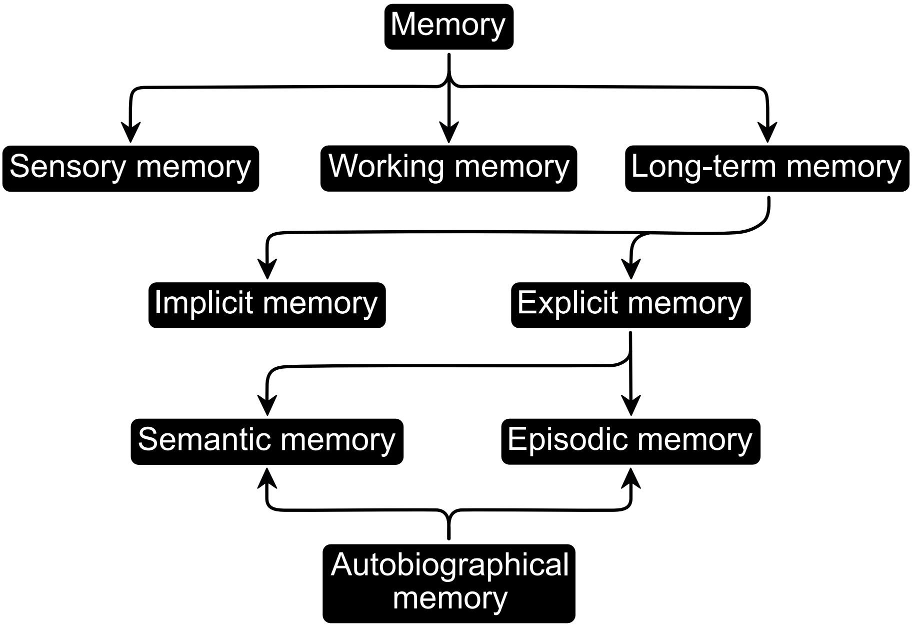


Figure 1: An overview of the different subcategories that memory can be divided into.

can be aware of the fact that you are being conditioned to associate a certain sound with getting a piece of candy, this explicit knowledge will not be the reason why you later react positively to that sound. Given this definition, implicit memory contains a plethora of memory processes such as conditioning (which the candy scenario above is an example of), priming, habituation, and motor learning.

Explicit memory, on the other hand, refers to the type of memory that *do* requires cognizant processes. Here, a further division is often used, namely that between *semantic memory*, *episodic memory*, and *autobiographical memory*. Semantic memory is the knowledge about such things as facts, concepts, ideas, and words; basically explicit memories that are decoupled from when and how they were learned. For example, to know what a quokka is[10], one does not need to know how this fact was acquired.

In contrast, episodic memory, which will be the focus of this disser-

it myself in this context. What I here refer to with the word *cognizant* is simply that the person can report knowledge of the matter at hand.

[10] If you do not know, look it up right now!

tation, has been defined by Tulving (1972, 2002) as the compound memory that stores the *what*, *where*, and *when* of an event. An everyday example of this would be to remember a meeting with a group of people in terms of what happened, as well as placing all separate events into space and time. Episodic memory would here be involved in answering questions like “Where was this person when he said this and that?”, “What did she say?”, and “Did this event occur before that event?”. As such, testing episodic memory can involve approaching it from many different angles, using every imaginable modality. A few examples would be remembering images, routes, faces, and locations. However, the most common way to measure episodic memory is probably by asking the participants to remember words from an earlier read or heard word list. One standardized example of this is the *California Verbal Learning Test* (commonly abbreviated CVLT; Delis, Kramer, Kaplan, and Ober, 1987), where the subject, on multiple occasions, is asked to remember the contents of a shopping list that they continuously are given the chance to rehearse.

Finally, autobiographical memory is a form of memory about oneself (Conway, 2005; Conway & Pleydell-pearce, 2000). In this regard, it encompasses specific memories of events but also more general and abstract concepts, sorted into different themes, such as knowledge about what activities one liked as a kid. As such, autobiographical memory is a form of combination of semantic and episodic memory, and will therefore not be included or further explored in this thesis.

1.5 MEAN VS. VARIANCE SEX DIFFERENCES

Most, but not all, research that has been done on sex differences has first and foremost been about sex differences in means. Here, the researchers compute the average performance/behavior/characteristic for men and women separate and then compare these means. If these numbers then can be shown to be different according to some statistical method, sex differences are proclaimed.

While mean differences absolutely are useful in and of themselves, they only reveal part of the picture. For example, simply knowing that two groups differ does not automatically enable you to predict the distribution among them in the extremes, that is, among the top or bottom achievers. Here, the concept of sex differences in variance comes in handy. Basically, when investigating this concept, you compute how much variation there is *within* each group and then make a comparison between them to see how much they deviate from each other. Given that you know both the mean and variance sex differences, you can now predict the ratios of the sexes in the extremes of the distribution. This is, of course, dependent on the fact that both groups are more or less normally distributed, an assumption that, in general, surprisingly often is satisfied for different types of empirical data (Frank, 2009), including psychological factors.[11]

Going further, in order to keep things organized, I will first, in the very next section, focus on what has been found regarding sex differences in general, in cognition, and in episodic memory when it comes to the measured means of men and women for different traits and abilities. Then, in the section after that, I will turn to describe what has been found within these fields with regard to differences in in-group variance between men and women.

1.6 MEAN SEX DIFFERENCES

1.6.1 GENERAL

When it comes to sex, there are some major differences between men and women that are totally obvious, and for most individuals it is close to impossible *not* to directly categorize others into either males or females. Except for the unmistakable differences such as the different reproductive system and physical appearance, there are, for example, large differences in the average strength and length between the sexes (Fryar, Gu, Ogden, & Flegal, 2016).

Turning to psychological traits, one of the most obvious differences between men and women is probably that they have dramatically different preferences for sexual and romantic partners. That is, the overwhelming majority of people are heterosexual (Bailey et al., 2016). Further, there has been a plethora of research within psychology showing that there are sex differences for things as disperse as personality (Costa, Terracciano, & McCrae, 2001), aggression (Bettencourt & Miller, 1996), narcissism (Grijalva et al., 2015), sexuality (Petersen & Hyde, 2010), mental health (Rosenfield & Mouzon, 2013), economic

[11] While investigating things outside of this assumption indeed would be both interesting and probably useful, this is beyond the scope of this thesis.

negotiation outcomes (Mazei et al., 2015), scholastic achievement (D. Voyer & Voyer, 2014), and pain sensitivity (Fillingim, King, Ribeiro-Dasilva, Rahim-Williams, & Riley, 2009) just to name a few.

1.6.2 COGNITION

When it comes to mean sex differences in cognition, one of the most well-researched and well-confirmed patterns is that of a difference between verbal and spatial tasks. What has been found here is that, in general, women tend to do better at tasks that are more verbal, and men tend to do better on tasks that are more spatial. In a meta-analysis by Hyde and Linn (1988), the authors concluded that an advantage for women could be found for many verbal tasks but not for all. For example, while women excelled at verbal fluency and verbal production tasks, no differences were found when assessing vocabulary, and men even outperformed women on analogy solving. Recent, large-scale, international studies have generated similar results, where teenage girls are better than teenage boys at reading comprehension (Stoet & Geary, 2013) (with the advantage for females here being about three times as large as the advantage for males in mathematics, which is discussed further down) and where adult women perform better than adult men on verbal fluency tasks (Maylor et al., 2007), a type of test where the participants are supposed to generate as many words as possible beginning on a certain letter or belonging to a certain category.

Turning to spatial/visuospatial abilities, D. Voyer, Voyer, and Bryden (1995) showed in a meta-analysis that men tend to have an advantage over women here. However, just as with the verbal tasks presented above, depending on the exact type of spatial task performed, the overall pattern also differs. For example, the largest sex difference that was found in this meta-analysis was for mental rotation tasks. This difference was at its largest when the participants worked under time pressure and when the object was in 3D rather than in 2D (D. Voyer, 2011; D. Voyer et al., 1995). However, men's advantage was not as strong in tasks having to do with spatial perception and dropped even more for spatial visualization (D. Voyer et al., 1995). Also, mathematics might not be an ideal example of a spatial/visuospatial task, but to the extent that it is, sex differences favoring males have been found (Stoet & Geary, 2013).

1.6.3 EPISODIC MEMORY

In 1975, Maccoby and Jacklin conducted a large review of sex differences for many different psychological domains, including memory where they could not find an effect. However, this might have come about because several different memory types were intermixed, and later studies have suggested that there actually are differences between men and women here. For example, in a large study (Herlitz, Nilsson, & Bäckman, 1997) of 1000 adults, women performed better on episodic memory tasks where they were asked to remember activities, words, facts, or objects. Even though all these tasks are not verbal *per se*, they can all be supported by verbal processes. For example, even if objects are presented as images, they can be remembered as words as long as the images shown are not too abstract. In another study (Astur, Ortiz, & Sutherland, 1998) that investigated more spatial tasks, men outperformed women when navigating through a maze with almost no external cues. However, when the task changed so that verbal support could be used, much smaller effects were found and in some cases no sex differences at all. Further, studies investigating both more verbal and more spatial tasks in the same subjects have also been conducted, where the results have been similar to other type of cognitive tasks, meaning that women tend to perform better on more verbal oriented memory tasks while men tend to have an advantage in more spatial tasks(Herlitz, Airaksinen, & Nordström, 1999; Lewin, Wolgers, & Herlitz, 2001). This pattern, where there seem to be a sort of verbal-spatial spectrum that goes hand in hand with the relative performance of men and women is something that has been suggested in later reviews (Andreano & Cahill, 2009; Herlitz & Rehnman, 2008). It should be noted, though, that men only seem to outperform women when tasks rely heavily on spatial processing, which for example is highlighted by the fact that women seem to perform better than men when it comes to object location memory (D. Voyer, Postma, Brake, & Imperato-McGinley, 2007).

The hypothesis that the same verbal/spatial pattern that is present for sex differences in general cognition also should be present for episodic memory would make sense; since memory is not an isolated cognitive ability, working independently of other cognitive mechanisms, one should be able to draw from skills present for non memory related processes. For example, being able to effectively structure and visualize spatial material should influence the ability to also later recall that material. As an analogy, an expert chess player should easily be able to recall complex states of the chess board in a way that a novice probably could not.

As have been presented above, circumstantial evidence already exists regarding sex differences in episodic memory and how different factors might modify it. However, no large-scale investigation of this matter — which could confirm the verbal/spatial pattern hypothesized above as well as examine other aspects of how sex differences might vary — has been performed. Therefore, for Study I of this thesis, we conducted a large meta-analysis of this field.

1.7 VARIANCE SEX DIFFERENCES

1.7.1 GENERAL

As early as 1894, Ellis suggested that men as a group varied more than women in mental and physical characteristics, and since then investigations of large data sets across regions and time have confirmed and expanded on these findings. For example, men are more variable when it comes to such things as birth weight, blood parameters, and juvenile physical performance (Lehre, Lehre, Laake, & Danbolt, 2009). Further, in a recent study, not yet published, it was demonstrated that the same pattern is present for subcortical volumes, cortical surface areas, and cortical thickness and that these differences are seen throughout the lifespan (Wierenga et al., 2020).

1.7.2 COGNITION

In 1932, a sample of 87,498 Scottish schoolchildren born in 1921 (which constituted about 95% of all children in that cohort) had their intelligence assessed with a large test battery not designed to minimize sex differences (for Research in Education, 1933). When Deary (2003) later analyzed the data, he could show that boys were overrepresented both in the higher and lower tails of the distribution (see Figure 2; Deary, 2003), something that he concluded “in part [might] explain such cognitive outcomes as the slight excess of men achieving first class university degrees, and the excess of males with learning difficulties.” The same survey was repeated in 1947 (for Research in Education, 1949), where Johnson, Carothers, and Deary (2008) could confirm the same pattern. However, for both surveys, boys and girls were very similar in mean intelligence. Similar patterns have been found by, for example, Hedges and Nowell (1995) who investigated the performance of over 100,000 schoolchildren, tested between 1960-1992 in the United States, and found that males showed a larger variance in performance compared to females. One of the few exceptions here was associative memory (which I will come back to in section 1.7.3). In a similar vein, Strand, Deary, and Smith (2006) found that in a national sample of 320,000 UK 11-12 year-olds, boys were more variable on quantitative, and non-verbal reasoning, and that they were overrepresented in the 10% lowest-performing students on verbal skill. This result was replicated by Lohman and Lakin (2009) in the United States with an almost as big sample of school children of all ages, tested between 1984 and 2000, where the same abilities were assessed. Similarly, analyses of the *Programme for International Student Assessment* (PISA) studies have shown that boys consistently are overrepresented in the top and bottom performers in mathematics compared with the sex ratios that would be expected from the mean sex differences alone, suggesting that their test score variance is larger than what it is for girls (Stoet & Geary, 2015).

However, this pattern of larger variances in the male group is not always found. In addition to associative memory, Hedges and Nowell (1995) found no consistent variance sex differences in reading comprehension and perceptual speed. Further, when Irwing and Lynn (2005) conducted a meta-analysis of 22 studies from around the world on progressive matrices performance (a task where you have to guess the last part of a given pattern of abstract forms) in university students, they found that while men had a higher performance average, women were more variable on the standard version of the test compared to the more advanced where no variance differences were found. Also, in a large, population-based Romain sample, no variance sex differences could be found either for general intelligence or for more specific, second-level cognitive abilities (Iliescu, Ilie, Ispas, Dobrean, & Clinciu, 2016).

1.7.3 EPISODIC MEMORY

To my knowledge, there has been no other investigation of sex differences in variance for episodic memory except for the study by Hedges and Nowell (1995) mentioned in the previous section, where associa-

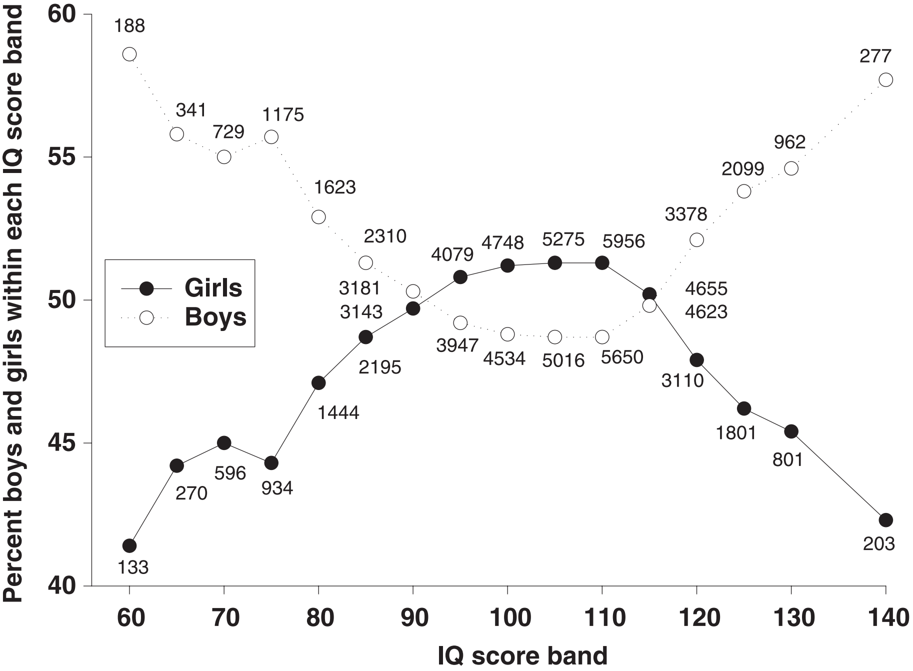


Figure 2: Figure 1 from Deary (2003), showing the distribution of IQ scores in a large sample of Scottish 11-year-olds tested in 1932. Here, both the number of people as well as the total percentage of all subjects is shown separately for the two groups within 5-points wide IQ bands.

tive memory in three national U.S. surveys — in total comprising over 100,000 school children — was investigated. Here, results showed no clear pattern, with one sample showing larger variance for boys, one showing larger variance for girls, and one showing neither. Since the result of this investigation in variance differences in itself was highly variable, and since only one specific type of memory task was investigated, this means that more research is needed in order to say something regarding this matter. This will be the focus of Study II of this thesis.

1.8 EXPLANATIONS OF SEX DIFFERENCES

So far, I have mostly described what sex differences there are without really touching on any explanations for *why* these sex differences also exist, a question I will now turn to. Here, it is important to bear in mind that what constitutes an explanation depends on what perspective you are interested in. For example, you could propose why something should be the case by referring to how adaptive it has been during the course of evolution. However, this does not mean that you now understand *how* it comes about any more than understanding *why* humans build bridges helps you understand *how* bridges are built. Similarly, you could put forth an explanation that a cognitive function activates a certain part of the brain more than another, but without a grander perspective of the ins and outs of how the brain works and give rise to human behavior this has limited explanatory power (see Jonas and Kording, 2017).

Given the reasoning above, the best way to understand why sex differences might arise is to approach the question from several different point of views. For the scope of this thesis, I will try to make a contribution regarding one of these perspectives, namely a societal explanation. However, I will start by going through both possibly evolutionary and biological explanations.

1.8.1 EVOLUTIONARY EXPLANATIONS

I will call something an *evolutionary explanation* if the way things are can to some extent be derived back to what has been beneficial for the spreading of certain genes that controls the phenotypic expression in question. From a theoretical point of view, major sex differences should only appear in a species if males and females have different roads to achieve reproductive success (which, evolutionarily speaking, is the only thing that matters), something that in turn only can be the case if the sexes' differ in what roles they play in the reproductive process (which they do since this is what defines sex in the first place). A clear example of this would be something that already has been touched upon in section 1.6.1, namely that most people are heterosexual (Bailey et al., 2016), which in turn means that men and women are attracted to dramatically different attributes in other humans. This can simply be derived back to the fact that historically, those individuals that were more attracted to the same sex did not procreate to the same extent as those that were more attracted to the opposite sex, in turn leading to genes controlling for homosexuality being less prevalent in the gene pool. However, this explanation breaks down at a certain point, and there are a number of theories regarding why, in an overwhelmingly heterosexual environment, it might be evolutionarily beneficial to be or lean towards homo- or bisexuality (Savolainen & Hodgson, 2016). For example, one such theory suggests that male homosexuality can be beneficial for an individual with an older brother, since it decreases competition for resources necessary for mating, thereby giving the sibling a sort of *carte blanche* to finding one or several suitable mates, something that will benefit the homosexual individual as well since he shares, on average, half his genes with his brother, which in turn will be passed on to the brother's children (Apostolou, 2013).

Moving one, one of the most important differences between males and females when it comes to the reproductive process is that males, in theory, can have an almost infinite number of offspring while females are restricted by the number of pregnancies[12] they can fit during their lifetime.[13] Further, a consequence of this is that in order for a male to spread his genes through means of having a child, the theoretically lowest amount of resources that he would have to devote would be the time and energy it takes to have sex with a female. He could then move

[12] In order to not have this section being longer than the rest of the thesis combined, I am going to restrict myself to only speak about mammals. So no egg-laying, sequential hermaphroditic, male pregnancy activities allowed!

[13] As a concrete example, it has been suggested that about 0.5% of all males in the world are ancestors to Genghis Khan, the emperor of the Mongol Empire who lived less than a thousand years ago (Zerjal et al., 2003), something which probably was achieved by him having several orders of magnitude more children than what a single woman ever could have.

on, and if the female does not make it on her own, his losses would, evolutionary speaking, be minimal given his investment. On the other extreme, he could plow all his resources into the female during pregnancy and the child postnatally. While both of these pathways could be successful strategies, their feasibility would depend on the male's other traits and qualities, as well as the environment in which he operates. That is, the strategy of impregnating a lot of females would probably not be especially effective for an unattractive male who will be given limited opportunities for coitus. Neither would it be a good strategy if there were a real shortage of other females to impregnate. On the other hand, in order for a female to have a child, she will *at minimum* have to go through a long, very resource-demanding pregnancy (and in the case of humans, historically a rather high risk that either she, her offspring, or both perish during childbirth), during which she will not be able to reproduce with someone else. This immediately narrows her options for successful strategies compared to a male. For example, abandoning the child would be a sort of evolutionary chicken race where the mother almost always would have more to lose from it than the father. Also, having sex with any random, willing male would for a time block her from later having children with a more suitable male. The point here is that from an evolutionary perspective, males have a large spectrum of strategies to draw from when it comes to successfully spreading their genes through means of procreation.

When it comes to variance sex differences, as already have been covered in section 1.7, males tend to be more variable than females. One theory that has been put forth to explain this pattern links it to the different strategies for procreation outlined above (Moore, 1991). According to the theory, because of the larger set of viable strategies available for males compared to females, there would also be a larger set of different phenotypes present in the male group compared to the female group. This would, in turn, translate into larger variance for different traits and skills for males as a group. [14]

Depending on different circumstances, species can evolve toward being more or less monogamous/polygamous (Kleiman, 1977). For example, if partners are scarce, it might pay off for males to stay and guard their

[14] A brief version of this explanation had to be removed from Study II in order for the paper to be accepted because it was judged by the journal to potentially drive a sexual stereotype.

mate against other suitors rather than to look for new mates (Schacht & Bell, 2016). Here, it has been shown that monogamous species exhibit less sex differences than polygamous species because their roles will more overlap in the former scenario (Gaulin, 1992). For example, while it may pay off for a polygamous male to have a large body size — something that ultimately might let him compete with other males and impregnate a lot of females — if he is monogamous, the cost of having a large body size makes less sense since the potential payoff is smaller.

The same pattern described in section 1.6.2, that men generally perform better on spatial tasks than women, has also been seen in many animals such as mice, rats, voles and monkeys (Gaulin, 1992; Jacobs, Gaulin, Sherry, & Hoffman, 1990; Jones, Braithwaite, & Healy, 2003). There are a number of different evolutionary theories regarding why males are better at spatial tasks where the one with the strongest support relates it to range size. More specifically, males in polygamous species usually have larger territories than the females whereas these two tend to overlap in monogamous species. This is a consequence of the polygamous males having a strategy to find and mate with many different females as opposed to just one as is the case with the monogamous males. Here, it has been seen that in species where males have larger range size than females, they also perform better than the females on spatial tasks, something that is not seen in species where the sexes have more equal territorial sizes. It is hypothesized that the underlying reason for this relationship is that males with larger range sizes than females also need better spatial skills, since navigating around in the environment is something that they need to be more proficient at. Since humans are not strictly monogamous creatures, this could then, at least partly, explain why men perform better than women on spatial tasks. One could, in a similar way as above, speculate about how women might be more benefited by verbal skills than men. However, since only humans display verbal skills, there are no animal studies regarding this matter. An evolutionary explanations would therefore be more of a just-so story, something that can be useful but still has less scientific proof than explanations like the ones for sex differences in spatial cognition presented above (Holcomb, 1996).

1.8.2 BIOLOGICAL EXPLANATIONS

I will call something a *biological explanation* if the state of things can be related to the constitution of some biological traits, such as the composition of the genes or the structure of the brain. On the genetic level, what differs between men and women is that for the 23rd chromosome pair (the so-called sex chromosomes) women have an X chromosome while men have a Y chromosome, meaning that men in total have fewer base pairs of genetic code than women. Even though these genetic differences alone should give rise to some phenotypic disparaties between men and women, most sex differences probably comes about because of gene/hormone interactions (Gaulin, 1992), meaning that autosomal genes (that is, genes not located on any of the sex chromosomes) are expressed differently depending on the hormone environment, which in turn often is a stable indicator of the sex of the individual. It is unclear exactly to what extent psychological sex differences arises because of this mechanism, but some evidence points to, for example, spatial abilities being affected by it (Berenbaum & Beltz, 2016).

Focusing on the genetic differences alone (that is, without the influence of hormones), one theory (Reinhold & Engqvist, 2013) has tied this to the fact that men seem to be more variable than women (see section 1.7). That is, for the overwhelming majority of all genes, their phenotypic expressions are influenced by two versions of the gene, one on each chromosome. This means, for example, that mutations of one copy of the gene in many cases can be counteracted by the other copy. However, for men, when it comes to genes on the sex chromosomes, some of these only have a single version since there are some genes "missing" on the Y chromosome. This has the effect that mutations on these base pairs more often actually are expressed for men, which in turn means that they show a larger variance than women. In addition, it should be said that for some species, females have the XY chromosome pair, meaning that *they* should be more variable instead, something that, for example, has been shown when it comes to body size (Reinhold & Engqvist, 2013).

Turning to the brain itself, if there are differences on the group level between men and women when it comes to behavior and cognitive performance, as have been shown in section 1.6 and 1.7, these differences *have* to be manifested in the nervous system somehow. [15] However,

[15] This logic holds even for mind-body interaction dualist like myself as long as the

one could pose the question of to what extent the brain differences that have been detected also are linked to behavioral and performance differences, and if so in what way (Grabowska, 2016; McCarthy, 2016). That is, even if observed discrepancies in behavior and performance *have* to have brain-based explanations, it is still possible that other brain differences than the ones that already have been found are responsible.

When it comes to biological differences in the brain, it has been confirmed that there are differences in neuroanatomy, neurochemistry, physiology, and connectivity (Andreano & Cahill, 2009; Ingalhalikar et al., 2014; Tian, Wang, Yan, & He, 2011). Men and women also tend to differ in many cases with respect to how the brain processes information. For example, sex differences have been found in how men and women process faces (Proverbio, 2017), as well as emotional expressions in others (Kret & De Gelder, 2012).

Further, observable brain differences do not automatically have to lead to observable behavioral differences, one possible reason being that the behavioral consequences simply are too small to detect. However, the same type of behaviors could also be manifested through different brain processes. Haier, Jung, Yeo, Head, and Alkire (2005) have for example shown that IQ is correlated with activities in different areas of the brain for men and women. While the frontal lobe is important for both sexes, men seem to rely more on the parietal lobe and women on Broca's area. Further, Jordan, Wüstenberg, Heinze, Peters, and Jäncke (2002) showed that men and women who are equally good at mental rotation tasks nonetheless showed differences in brain processing. The take-home message here is that when studying the brain, one has to be careful drawing conclusions about behavior.

Differences in the brain between men and women have also been found when it comes to episodic memory. For example, Young, Bellgowan, Bodurka, and Drevets (2013) found several brain discrepancies between the sexes for autobiographical memory irrespective of the valence of the content, something that they interpreted to reflect sex-specific cognitive recall strategies. Further, on a test of two different episodic memory tasks — more specifically *Wechsler Memory Scale - Revised* (commonly abbreviated WMS-R) and the California Verbal Learn-

belief is held that the mind to at least some extent relies on the brain to carry out behavior and perform computations, a position that just about every scientist and philosopher probably would agree with.

ing Test (already described in section 1.4) — Ragland, Coleman, Gur, Glahn, and Gur (2000) found that performance for women, but not men, was correlated to activity in the temporal pole.

1.8.3 SOCIAL EXPLANATIONS

I will call something a *social explanation* if it comes about because of the social environment that a person grows up and/or live in.[16] As such, there are many different type of potential social explanations that could be brought up when it comes to sex differences, for example, how the behavior of families, friends, and strangers might affect the sexes differently, or how stereotypes and gender roles might have an influence. However, for this thesis, I will focus on grander, more societal explanations, and this is also what we will investigate further in Study III.

The *Flynn effect*, first described by James Flynn in 1987, is a term that describes the phenomena that has been observed for over 100 years where the IQ scores steadily have gone up over time all across the globe (Williams, 2013). Some factors that have been suggested as explanations for this effect are increased education, gradually smaller family sizes, and improved health care (Pietschnig & Voracek, 2015). This is supported by the fact that the Flynn effect is largest in parts of the world that have experienced the most dramatic increase in living conditions (see for example Weber, Dekhtyar, and Herlitz, 2017).

Some researchers have proposed that there is a link between improvements in living conditions, like the ones mentioned above, and increases in women's cognitive performance when compared to men's, in a way suggesting that the Flynn effect and similar phenomena are affecting the sexes differently. One explanation for why this might be the case could be that in underdeveloped regions, women tend to have less access to social goods, for example education, than men. Possibly then, when living conditions increase in general, women may benefit more. That is, imagine a purely illustrative example where the average boy in a certain country goes from spending five years in school to ten while the average girl goes from getting no education at all to spending

[16] However, it is important to note that explanations like these never can be fully decoupled from evolutionary explanations, the likes I went through in section 1.8.1. In the words of Gaulin (1992): "Any claim that 'x is the result of experience' must explain why experience has that particular effect rather than any of an infinite array of possible effects. Such explanations can only be evolutionary."

five years in school. Even though both sexes are given five extra years in school, it stands to reason that girls should experience a larger improvement than boys in for example mathematical ability (even if boys still perform better in absolute terms) because of diminishing returns that might affect boys to a larger degree.

As a concrete example of this pattern, Guiso, Monte, Sapienza, and Zingales (2008) investigated the 2003 cross-cultural and cross-sectional PISA survey of mathematic competence in 15-year-olds found that while boys generally perform at a higher level then girls, in countries with relatively high gender equality, this advantage for boys was smaller. Conversely, in these countries, the advantage for girls in reading comprehension was further enlarged. Finally, it could also be seen that the difference between mathematical ability and reading ability shrunk for girls when gender equality increased, something that was not the case for boys. Else-Quest, Hyde, and Linn (2010) extended these findings by investigating the 2003 PISA data together with data of 15-year olds from the 2003 cohort of the *Trends in International Mathematics and Science Study* (TIMSS). Again, results showed that females' relative mathematical performance in comparison to males was positively associated with increases in gender equality. Further, Weber, Skirbekk, Freund, and Herlitz (2014) carried out an investigation of the *Survey of Health, Aging and Retirement in Europe* (SHARE) data, in which the subjects are older than 50 years and live in Europe. Here, yet again, results showed that improved living conditions and more gender-inclusive education were related to relative improvements in women's performance in comparison to men for verbal episodic memory (which I will come back to shortly), numeracy, and category fluency.

However, Stoet and Geary (2013, 2015) have questioned the generalizability of some of these results, claiming that the PISA data only show these patterns for the 2003 cohort, while the cohorts from 2000, 2006, and 2009 lacked this relationship. Further, Lippa, Collaer, and Peters (2010) studied line angle judgment and mental rotation performance in 18-40 year-olds from 53 different countries tested online and found that men not only performed better than women in all counties tested, but also that these sex differences were greater in men's favor in countries with more gender equality, something that runs contrary to the pattern suggested above.

Turning to episodic memory, the overall pattern has been more consistent. Here, two large scale investigations have been performed. First, as already mentioned above, when Weber et al. (2014) investigated data from the SHARE study, they concluded that when comparing with men, women's relative performance on episodic memory increased with better living conditions. Second, Bonsang, Skirbekk, and Staudinger (2017) investigated how sex differences in verbal episodic memory performance were moderated by gender equality values in society across both regions and age groups. More specifically, performance data on episodic memory from the *U.S. Health and Retirement Study*, the *Survey of Health, Ageing and Retirement in Europe*, the *English Longitudinal Study of Ageing*, and the *World Health Organization Study on Global AGEing and Adult Health* was used in conjunction with participant judgments from the *World Values Surveys* regarding if they agreed with the statement "when jobs are scarce, men should have more right to a job than women" as well as the number of protestants in each country (which was interpreted as a proxy of conservatism on gender issues). Results showed that both of these measures were linked to women's relative improvement in verbal episodic memory compared to men's (see Figure 3).

Importantly, both studies presented above (Bonsang et al., 2017; Weber et al., 2014) were performed on datasets where participants were above 50 years of age. While the data collections were not spread out in time, the authors of both studies used different values for the predictor variables for each age group in each country. For example, when Weber et al. (2014) investigated GDP per capita for a certain age group from a certain country, they used the GDP per capita value from the year when the participants in that age group were about 25 years old. This was then combined with other measures, handled similarly, to derive a regional development index for each specific age group in each specific country. One way that the results from these studies could be expanded upon would be to incorporate data from even more countries, collected at many different time points instead of a very narrow time frame. This is what will be done in Study III of this thesis.

1.9 OVERARCHING RESEARCH QUESTIONS

For this thesis, there are three overarching research questions/themes that will be investigated through three different studies:

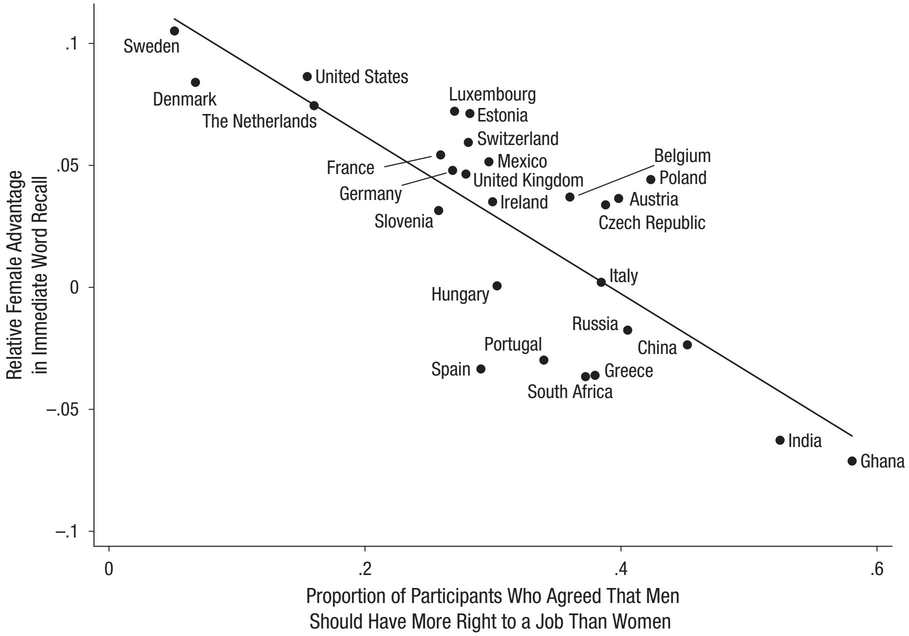


Figure 3: Figure 1 from Bonsang et al. (2017) showing, for each country, the association between relative female advantage (compared to men) in word recall and the proportion of people over 50 years in age agreeing with the statement: “When jobs are scarce, men should have more right to a job than women”.

1. Do men and women differ at the group level when it comes to mean differences in episodic memory performance and, if so, what factors influence this difference? This will be investigated in Study I.
2. Do men and women differ at the group level when it comes to variance differences in episodic memory performance and, if so, what factors influence this difference? This will be investigated in Study II.
3. Are men and women affected differently by societal progression when it comes to mean differences in episodic memory performance? This will be investigated in Study III.

2 METHOD

2.1 DATASET

In order to investigate the research questions outlined in section 1.9, we set out to create a dataset of relevant studies. More specifically, our goal was to capture as much as possible of the research that had been conducted on episodic memory where both men and women were tested.

2.1.1 DATA COLLECTION

Two different database queries were performed (see Figure 4 for an overview of the whole data collection phase). The first one was performed in *PsycINFO* and *Medline* and spanned from January 1972 to September 2001 where the search terms “memory” and “sex OR gender” were used. The term *episodic memory* has not been used consistently in the literature, which is why the broader concept *memory* was used instead. Also, not adding the “sex OR gender” would surely result in more relevant studies showing up, but this search resulted in over 65,000 abstracts, which was practically infeasible to go through. Instead, the final search resulted in 2,425 abstracts.

The original intention was that Study I would be completed after this first search, a plan that never was realized. Instead, when enough time had passed since the first search, an additional database query had to be

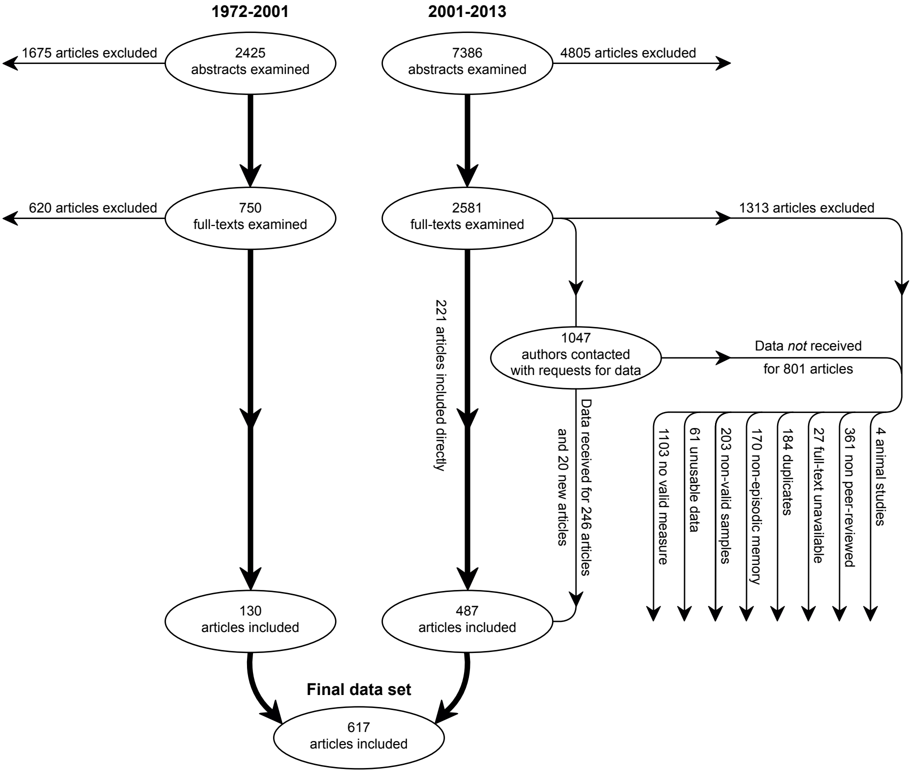


Figure 4: Figure 1 from Asperholm, Högman, et al. (2019) depicting the data collection phase of the dataset that constituted the basis of the studies in this thesis.

performed to get the dataset up to date. This database search was performed in *PubMed* and *PsycINFO*, spanning from September 2001 to 25 November 2013. Here, the search terms "memory", "sex OR gender", and "humans" were used, the last term added to even further narrow the results. This resulted in 7,386 abstracts.

After retrieving all abstracts, the next step was to go through them and, if deemed interesting, retrieve the full text to investigate them further. For the first search, full-text articles were retrieved if it somehow was indicated in the abstract that males and females had been compared or if this could be reasonably suspected. For the second search, this process was even more inclusive where full-text articles were retrieved if it could not be ruled out from reading the abstract that both males and females had performed some kind of episodic memory task.

There were four main criteria that each study/sample/measure needed to satisfy in order to be included in the final dataset:

1. The study had to be published in a scientific, peer-reviewed journal (meaning for example that book chapters and dissertations were not included), and the data had to be unique for that particular study (meaning that if several articles used the same dataset, we kept the one deriving the outcome on the largest amount of participants).

2. The participants that partook could not have been selected based on some type of illness, disease or disorder, and they could not have been subjugated to any type of experimental manipulation that seriously may have affected their memory performance. This meant that while, for example, different type of health conditions could be represented in the dataset, its prevalence in the sample should more or less reflect the prevalence of that condition in the general population.

3. All participants had to be exposed to the same material during the encoding phase. For example, questions of what you ate for breakfast would not qualify since different participants would have eaten different meals (or maybe even nothing), meaning that the difficulty of the task would vary between them (not to mention that it is impossibe for the experimenter to know the correct answers).

4. A direct measure of episodic memory accuracy, such as the num-

ber of words remembered from a previously seen word list, had to be reported. This means that, for example, total response time would not qualify since this would measure how fast, rather than how accurate, someone remembered something.

All articles that had been selected for inclusion in the first data collection phase were reexamined and recoded in the second data collection phase in order to achieve a high level of consistency between the two searches.

In many cases, the researchers had no intentions of investigating differences between the sexes but rather to examine episodic memory with regard to some other aspect. Therefore, for the second data collection phase, if it was suspected that episodic memory performance data for both males and females existed, and if the article was not published more than ten years prior, a couple of requests to get the relevant data was sent out to the authors via e-mail. In total, 1047 authors were contacted, which resulted in us receiving unpublished data for 246 articles that we could use. In addition, we could sometimes be redirected to other articles where the relevant data had been published. Finally, sometimes the dataset used in a certain article came from a large, open database. Whenever this was the case, we went to the original source, something that resulted in four large databases being included in the study. These databases were: *Survey of Health, Ageing and Retirement in Europe* (SHARE; Börsch-Supan et al., 2013), *The English Longitudinal Study of Ageing* (ELSA; Steptoe, Breeze, Banks, and Nazroo, 2012), *Health and Retirement Study* (HRS; Sonnega et al., 2014), and *Study of Global Ageing and Adult Health* (SAGE; Kowal et al., 2012).

All in all, 617 studies were included in the final dataset, constituting 4,171 individual effect sizes, 1,370 independent samples, and 1,233,921 participants (with 564,433 males and 669,488 females). The full or a partial set was then used for all studies that we conducted in the context of this thesis.

2.1.2 HIERARCHICAL STRUCTURE OF THE DATA

The final dataset had a certain hierarchical structure to it: Every study could be made up of several different samples. For each of these samples (for example young and old participants), participants could have performed several different tasks (for example remembering words and remembering images), and for each of these tasks, several different outcomes could be measured (for example free and cued recall). This means that a study could consist of just one effect size (if just one measure was taken from one task for one sample) or many. For a visual illustration of this structure, see Figure 5, and for a discussion on how this structure needs to be handled statistically, see section 4.3 (the upshot of that discussion being that the standard meta-analytic procedure should not be used, but rather a model specifically tailored to deal with dependency structures like these).

2.1.3 MODERATORS

In addition to the data needed to compute an effect size (see section 2.2.2), a lot of other variables were recorded for each data point that then functioned as the basis for several different moderator analyses and subset divisions in the different studies. Most of these variables were only analyzed in Study I, but some of them were used throughout all of the studies. The most used variable, that in one way or the other was used in all three studies, was *type of material* to be remembered. This division consisted of eight separate levels, ranging from highly verbal to highly spatial, although for three of the levels (*faces*, *sensory*, *remaining*) this scale was not applicable. More specifically, these categories were:

Verbal. Words, sentences, facts, and conversations.

Images. Images of real or abstract objects and scenes.

Movies. Movie clips with or without sound.

Locations. Locations of objects. These tasks usually took the form of either having the participants placing objects where they previously had been or having them indicate what object previously had been at a certain spot.

Routes. Routes through space. These tasks could, for example, be about memorizing a path on a map or remembering the way through a 3D maze.

Faces. Images of human faces.

Sensory. Odors, tastes, and colors.

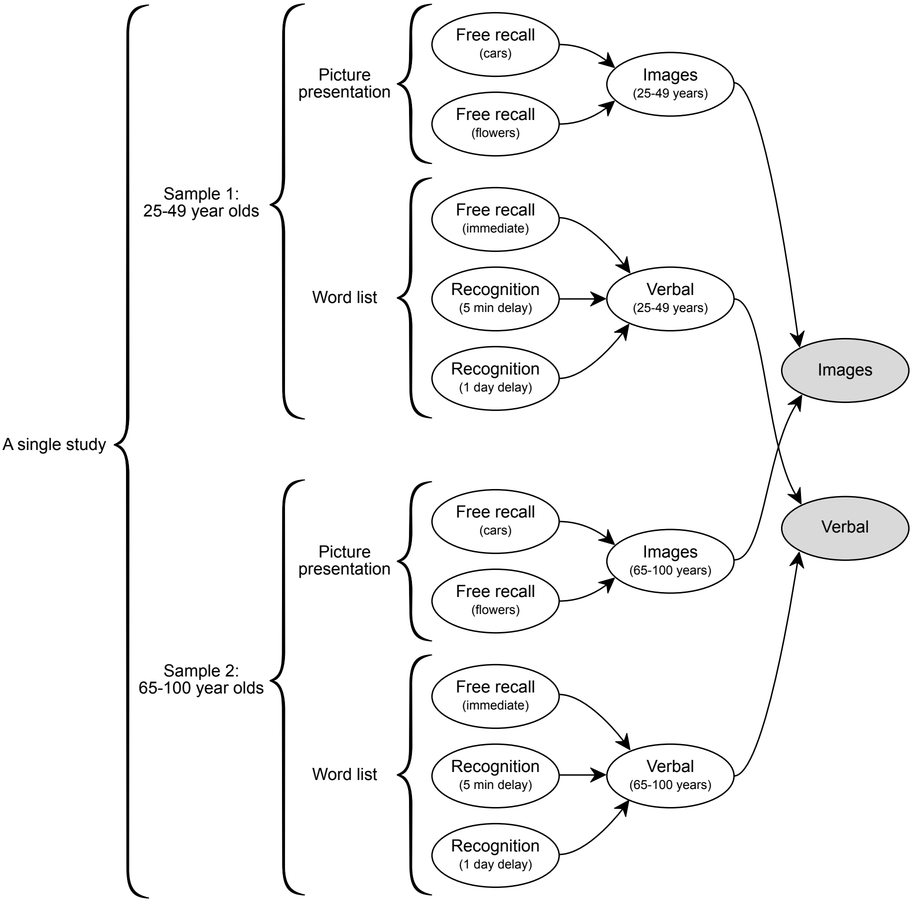


Figure 5: Figure S1 from Study I depicting the hierarchical structure of the data described in section 2.1.2. Here, a specific example is shown where two samples have been tested on two different tasks with several outcomes each. It is then showed how these different data points contribute toward a certain moderator, more specifically different levels of what type of material that was supposed to be remembered (see section 2.1.3).

Remaining. Material that could not be placed within one of the above categories, such as composite measures based on several of them, or simply tests that were not described in enough detail to make a confident judgment about which category to place them in.

Another set of variables that were used both in Study I and Study II pertained to possible bias in the dataset. More specifically, the following four variables were defined:

Database search. The distinction brought up in section 2.1.1 between studies gathered in the 1972–2001 or 2001–2013 database query.

Data source. The distinction brought up in section 2.1.1 between data that were recorded directly from published articles and unpublished data that were retrieved from authors via e-mail.

Study objective. The distinction between whether the goal of the study was to investigate sex differences or not (which was most often the case).

Sampling of subjects. The distinction between whether the sample used for a particular study was convenience based (which most often was the case) or if it was population-based, meaning that the authors made a serious effort to have the sample accurately represent the population they were investigating.

Finally, yet two more continuous variables that were used both in Study I and Study II were defined the following way:

Age. A continuous value representing the sample age. Whenever a mean age was given, this was also used as the *age* value of the sample. However, sometimes only an age range was given. In this case, the middle value of that age range was used as the *age* value of the sample.

Year. The publication year of the study from which the data point was taken.

2.2 STATISTICS

2.2.1 BASICS OF META-ANALYSES

For all studies included in this dissertation, the meta-analytic method was used for the main analyses, although it is important to note that this does not automatically make the studies into traditional meta-analyses. Rather, as will be seen, the meta-analytic method can — or even *should* — be applied to any type of data where the individual data points have some known variance, preferably with some kind of measure that can be used to weight the different data points according to how precise these variances can be assumed to be (in our case, that would be the number of participants in a sample). This can be contrasted to most common statistical analyses such as a simple t-test or an ANOVA, where the dependent variable usually only is a set of numbers that represent individual measuring points (meaning that they have no inherent variance).

So, when you have aggregate data as your raw data, meaning that you have both the mean and the variance for each data point, you can use this variance to weight each data point in the final estimate. Even better, if you know how many individual measuring data points each aggregate data point is made up of (in our case, the number of participants), you can compute a weighted variance based on how precise the data point in question can be assumed to be. That is, if you in your dataset, consisting of aggregated variables, would disregard the weighted variance of individual data points, you would also end up not knowing whether one aggregate data point comes from sampling something 10 or 10,000 times, information that you would like to have in order to weight them properly. So, for example, if you had aggregated data from a study with 10 participants and from a study with 10,000 participants where the means were different but the reported variances were the same, disregarding sample sizes, you would have to conclude that the actual effect size lies somewhere between the two means when it, of course, would be more sensible to conclude that the true effect size lies much closer to the 10,000 participants study than the 10 participants study.

2.2.2 EFFECT SIZES

There are two different types of effect sizes that will be heavily used as the dependent variables in the three studies of this thesis:

> **Cohen's *d*/Hedges' *g*.** The most common effect size that historically has been used in meta-analyses is *Cohen's d*, which is defined as
>
> $$\frac{m_{women} - m_{men}}{sd_{total}}$$
>
> where m_x stands for the mean of group x and sd_{total} stands for the pooled standard deviation. In other words, it is a measure of how much a group deviates from another in term of their pooled standard deviation, where the pooled standard deviation in turn is the combined standard deviation of the both groups where the assumption is that their underlying variances are the same.
>
> There are several different ways that the pooled standard deviation can, and have been, computed and depending on which method that is being used, researchers have in a quite inconsistent way (as unfortunately often is the case with statistical concepts and which I do not, regrettably, help to alleviate in this thesis) used concepts such as Cohen's *d*, Hedges' *g*, and Hedges' g* (Durlak, 2009; Grissom & Kim, 2005; Mcgrath & Meyer, 2006; Zakzanis, 2001). For this thesis, whenever Cohen's *d* is used, which it is in Study III, the pooled standard deviation has been computed using
>
> $$\sqrt{\frac{(n_1 - 1)sd_1^2 + (n_2 - 1)sd_2^2}{n_1 + n_2 - 2}}$$
>
> which is the way the statistical package *compute.es* (Re, 2013) implements it. Here, n_1 and n_2 stands for the number of measuring points for each group, or more specifically, in our case, the number of participants. Further, one common addition is to multiply the effect size with a correction factor
>
> $$1 - \frac{3}{4(n_1 + n_2 - 2) - 1}$$

The point of this correction factor — which is a computationally lightweight approximation of the more complicated, original equation (Hedges & Olkin, 1985, p. 104) — is to counteract upward bias in small samples (<20 participants). When using this correction factor, which is the case for Study I and II, we call the measure Hedges' *g*. This is, again, the way it is handled and called in the *compute.es* package (Re, 2013).

As mentioned in the previous section, in order to use an effect size measure in the meta-analytical method, you also need the *weighted variance* for that very measure, that is, the variance of the aggregated measure weighted according to the number of measuring points (or number of participants). To derive the variance for Cohen's *d*, we used the formula

$$\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}$$

and for the variance of Hedges' *g*, we simply multiplied the variance of Cohen's *d* by the squared correction factor (Re, 2013).

Throughout this whole thesis, positive values of *d* indicate that women perform better than men and negative values indicate that men perform better than women. The closer the value is to zero, the smaller the difference is. For example, a Cohen's *d* of 0.20 indicates that 58% of all women performed better than the average man. Corresponding percentages for d=0.30 and d=0.50 are 62% and 69%, respectively.

lnVR. The *variance ratio* (*VR*) is a measure of the difference in variance between two groups. It is computed by taking the variance of one group over the variance of the other, hence the name variance *ratio*. This then leads to that a *VR* of 2 means that the variance of one group is twice as large as another. However, one problem with this measure, when using it as the basis for different computations, is that when switching the numerator and denominator, the result is not the same distance away from 1 (the value given when the groups are equal) as it previously were (that is, $\frac{2}{1} \neq \frac{1}{2}$). This means that in a *VR* dataset where there are no overall variance differences present, such as $\{\frac{2}{1}, \frac{1}{2}\}$, the mean is not 1 but rather

$$\frac{(\frac{2}{1} = 2) + (\frac{1}{2} = 0.5)}{2} = 1.25$$

which means that another measure will have to be used for situations like these. Here, one can instead use *the natural logarithm of the variance ratio* (*lnVR*; Feingold, 1992), which has the advantage that it does *not* matter if you switch the numerator and denominator around, that is, $|\ln \frac{2}{1}| = |\ln \frac{1}{2}|$. The vertical lines here indicate absolute values (that is, values where you remove minus signs), which are necessary for this explanation since the point indicating no variance differences for *lnVR* is 0.

More specifically, we used

$$ln\frac{sd_1}{sd_2} + \frac{1}{2(n_1 - 1)} - \frac{1}{2(n_2 - 1)}$$

to compute *lnVR* and

$$\frac{1}{2(n_1 - 1)} + \frac{1}{2(n_2 - 1)}$$

to compute the variance of *lnVR* (Nakagawa et al., 2015) for Study II. Here, positive values mean that men have larger variance than women.[17]

2.2.3 SPECIFIC META-ANALYSES USED

Basically all analyses conducted for the three studies included in this dissertation were multi-level meta-analyses, conducted either with or without discrete or continuous moderators. Multi-level meta-analysis was chosen as the preferred model, rather than the random effects meta-analysis that is the more standard method, because this is one of the ways of handling the hierarchical dependency structure in the data described in section 2.1.2 (for more information about this method and

[17] The reason why, maybe somewhat confusingly, positive Cohen's *d*/Hedges' *g* values mean that women performed better while positive *lnVR* values mean that men were more variable is that positive values were used to indicate the most often expected direction of the measure in accordance with the original hypotheses of the studies in this thesis.

the arguments for using it, see section 4.3). For each data point, Cohen's *d*/Hedges' *g*/*lnVR* along with the variance of Cohen's *d*/Hedges' *g*/*lnVR* were used as the dependent variables.

In order to better understand the results of the three studies presented in the results section of this thesis (section 3), it is important to understand the different types of analyses that have been performed. Basically, there are three different meta-analyses that have been conducted:

> **Without moderators.** When conducting a meta-analysis *without* any moderators, you are testing the null hypothesis that the estimated overall effect size is equal to zero. As such, from running this analysis, you get an overall effect size estimate as well as a *p*-value pertaining to this null hypothesis.
>
> **With a discrete moderator.** When conducting a meta-analysis with a *discrete* moderator, you are testing the null hypothesis that the level of the moderator has no effect on the resulting estimate. As such, from running this analysis, you get a *p*-value pertaining to this null hypothesis. In addition, you also get individual estimates for each level of the moderator along with *p*-values for the null hypotheses that they are zero.[18]
>
> **With continuous moderators.** When conducting a meta-analysis with one or several *continuous* moderators, a so-called *meta-regression*, you are testing the null hypothesis that the moderator(s) are not predicting the effect sizes. As such, from running this analysis, you get a *p*-value pertaining to this null hypothesis. In addition, you get, just as with standard regressions, an estimate of the intercept as well as estimates of the slopes for each moderator, together with *p*-values for the null hypotheses that the intercept and the slope(s) are different from zero.

One thing to keep in mind going forward is that the analyses described above only are designed to test for whether there are differences in the dependent variable between men and women, meaning that the results presented in turn only can provide evidence *for* sex differences. This has the consequence that non-significant results only mean that there

[18] Getting all these results really involve fitting *two* different models, one where each level of the moderator is tested against an arbitrarily chosen reference level and one where each level of the moderator is tested against a fixed effect size (in our case, 0).

is not enough power to detect a difference. Consequently, this leads to that non-significant results neither can be used to determine that there are *no* sex differences (which in principle never can be shown, given that the sex difference could be infinitesimal small), nor to determine that differences are small enough to be meaningless. In order to draw conclusions regarding the latter, one would have to construct a test where the alternative hypothesis is that the effect lies within a certain, sufficiently small, interval around zero (which has to be based on one's subjective judgment of what the smallest, meaningful effect size is).

3 RESULTS

In this section, I will present the three studies that form the basis of this thesis. They can all be found in their entirety at the very end, and what I present here are more condensed versions, even though my own thoughts and interpretations at certain places are more expanded upon than what they are in the original articles.

3.1 STUDY I

Study I is titled *What did you do yesterday? A meta-analysis of sex differences in Episodic Memory*. It was published in *Psychological Bulletin* in 2019 (Asperholm, Högman, et al., 2019).

3.1.1 INTRODUCTION

As already briefly outlined in section 1.9, the goal of Study I was to investigate mean differences between men and women when it comes to episodic memory performance, first and foremost with regard to the type of material to be remembered, and second with regard to other moderators that might affect the pattern within each of these material categories. Regarding the type of material, it was hypothesized that this aspect would not only affect the outcome but also that the pattern of sex differences in episodic memory would mirror the pattern that has been observed for other cognitive traits, namely that women perform better on more verbal tasks while men perform better on more spatial tasks (see section 1.6.2). Other moderators that were investigated all had in common that they either previously had been shown to differentiate various groups (not only for sex) and/or were interesting because of the conclusions that could be drawn if they would prove to be relevant.

3.1.2 METHOD

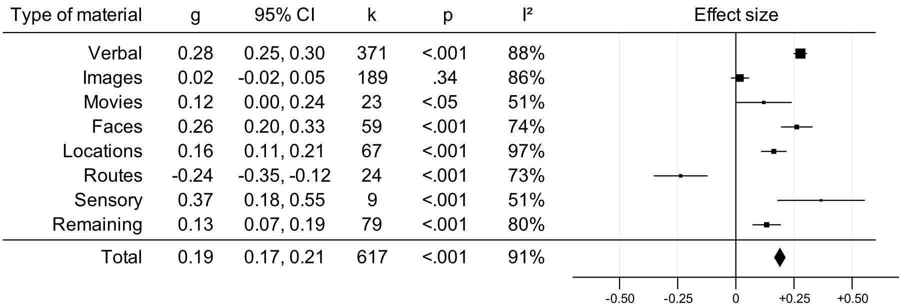
For this study, several five-level meta-analyses (further described and discussed in sections 2.2 and 4.3) were conducted on the full dataset (further described in section 2.1) as well as *material category* subsets (further described in section 2.1.3), using the dependent variables Hedges' *g* and the weighted variance of Hedges' *g* (further described in section 2.2.2) with or without a number of different moderators.

3.1.3 RESULTS AND DISCUSSION

First, a meta-analysis without any moderators was performed on the 617 articles contained in the full dataset. The result showed a significant advantage for females in episodic memory performance (g=0.19, 95% CI=[0.17,0.21], p<.001). Further, a meta-analysis with *type of material* to be remembered as a moderator showed that the sex difference in episodic memory performance, as hypothesized, was moderated by this aspect (see Figure 6). Also, the more specific, expected pattern where women would perform better for tasks relying on verbal processing while men would perform better for tasks relying heavily on spatial processing/abstract thinking somewhat emerged. It can be noted that for *images* no sex difference was found, and for *locations* women performed better than men.

Next, a number of moderator meta-analyses were performed at the task level, investigating each *type of material* subset separately as well as the *total* dataset. Here, the following moderators were investigated, which all had to do with aspects of the material to be remembered:

> **Nameable material.** Whether the material was nameable, meaning that it could be remembered using verbal, mental representations.
>
> **Emotional material.** Whether the material had a positive, negative, sexual, or neutral valence.
>
> **Paired material.** Whether the material was paired with some other material and the participants were supposed to remember all the individual associations.



Type of material	g	95% CI	k	p	I²
Verbal	0.28	0.25, 0.30	371	<.001	88%
Images	0.02	-0.02, 0.05	189	.34	86%
Movies	0.12	0.00, 0.24	23	<.05	51%
Faces	0.26	0.20, 0.33	59	<.001	74%
Locations	0.16	0.11, 0.21	67	<.001	97%
Routes	-0.24	-0.35, -0.12	24	<.001	73%
Sensory	0.37	0.18, 0.55	9	<.001	51%
Remaining	0.13	0.07, 0.19	79	<.001	80%
Total	0.19	0.17, 0.21	617	<.001	91%

Figure 6: Figure 2 from Study I showing how the sex differences in episodic memory performance varies with the type of material that one is supposed to remember (Omnibus *p*<.001; *I*^{2}=89%). Each row shows whether the effect size for that level of the moderator is reliably different from 0. The last row show the result from a meta-analysis without any moderators, performed on the full dataset. Positive values mean that women performed better than men. Explanation of headings: *g* = Hedges' *g*; 95% *CI* = the 95% confidence interval; *k* = number of studies; *p* = the *p*-value; *I*^{2} = statistic denoting the percentage of variation across studies that is due to heterogeneity rather than due to chance.

Here, it was only the *nameable material* variable that showed a somewhat consistent and interesting pattern (see Table 2 of Study I): For the *total* subset, women performed significantly better than men when the material was nameable (*g*=0.14; *p*<.001) and men performed significantly better than women when the material not was nameable (*g*=-0.22; *p*<.001). This was also the case for the *images* subset where females performed significantly better on nameable images (*g*=0.16; *p*<.001) while men performed significantly better on non-nameable images (*g*=-0.20; *p*<.01). The effects went in the same directions, although not significantly so, for the two remaining subsets where this division was applicable: *locations* (*k*=29; omnibus *p*<.20) and *routes* (*k*=19; omnibus *p*<.24). These resuts further strengthen the expected pattern with regard to the verbal/spatial continuum.

Next, another set of moderator meta-analyses were performed, again at the task level and again investigating each *type of material* subset separately as well as the *total* dataset. This time, the moderators related to aspects of how the material was learned or retrieved:

Repeated learning. Whether the participants were given several chances to learn the material or simply had to remember it after just going through it once.

Intentional learning. Whether the participants explicitly were asked to remember the material.

Retrieval support. Whether the participants were asked to freely remember the material, remember the material with the help of one or more cues, or whether they recognized the material later presented to them.

Delayed recall. Whether there was a delay between the presentation phase and the recall phase.

Results (see Table 3 of Study I) revealed no obvious and consistent patterns, but some interesting things can still be said: For *repeated learning*, the subsets *total*, *verbal*, and *images* showed significant omnibus tests, all going in the same direction with women increasing their performance more than men when repeating the material. However, the subset *Routes*, which had a barely significant omnibus test (while the *p* value is .05, it is rounded *down* to this value), showed the opposite direction. If this possible pattern reflects something real in the population, it might simply be that whenever a group outperforms another for a certain type of material to be remembered, this advantage is even further increased when given a chance to repeat the material. As an analogy, imagine two persons trying to learn a new language, one of them being good at learning new languages, the other one being bad at it. After one week, some differences between the two should be seen, although not that large. On the other hand, after one year, the first one should be speaking the language fluently while the other one still probably speaks at a beginner level.

For *retrieval support*, only the omnibus tests for the *total* and the *verbal* subset came out significant. Bear in mind that the overlap between these two sets always is very large though, in this case about 61% of the data points in the *total* set came from the *verbal* subset, so whenever one of them shows a certain pattern, there is a good chance that the other one will as well. For both of these sets, women performed significantly better than men for recognition, which should be the easiest task, even better for cued recall, and finally with the largest margin for free recall, which should be the hardest task. This suggests a pattern where the excelling sex outperform the other sex more and more, the harder the task is. However, as mentioned, none of the other subsets had significant omnibus tests, and from just eyeballing the estimated effect sizes, no similar pattern can really be seen.

Next, a number of moderator meta-analyses were performed at the sample level, again investigating each *type of material* subset separately as well as the *total* dataset. More specifically, in addition to *age*, *sampling of subjects*, and *year of publication* (that all were defined in section 2.1.3), *geographical region* was also investigated. This variable simply indicated where each study was carried out: Africa, Asia, Europe, North America, Oceania, or South America.

The results (see Table 4 of Study I) showed, first of all, that *geographical region* moderated the estimates within the two largest *type of material* sets: *verbal* and *total*. This pattern was further investigated in Study III. Further, *sampling of subjects* could, due to unbalanced data, only be investigated for the *verbal* and *total* sets, where the omnibus tests were significant; in both cases, the females' performance was even larger than males for convenience-based samples compared to population-based samples. It is hard to say why this is the case, but it should be noted that all four (very large) open databases that were included were categorized as population-based. If these somehow are different than the rest of the studies, they would most certainly impact the *sampling of subjects* category. Most notably, they all utilize similar type of verbal memory tasks. None of the meta-regressions for *year* came out significant, but for *age*, relationships were present for *verbal*, *images*, and *total* where both linear and quadratic regressions were fitted (see Figure 3 of Study I for scatterplots and fitted regression curves for this data). While the effect was negative and linear for *images*, the curve was parabolic for *verbal*, implying that gender differences were somewhat smaller for younger and older persons than for the rest.

Finally, possible bias in the dataset was evaluated through four different moderator meta-analyses (with the moderators being defined in section 2.1.3): *database search* (whether data was taken from the first or second database search), *data source* (whether data points were taken from publications or received from authors), and *study objective* (whether sex differences explicitly were investigated). The results (see Table 1 of Study I) showed no significant omnibus tests, meaning that no clear signs of bias in the dataset could be identified (for an extended discussion of this matter, see section 4.2).

In summary, I think that we, with this study, managed to give an exhaustive picture of what sex differences there are in episodic memory as well as which factors that are most important in order to understand the variation. More research can, of course, always be carried out, but I do not think that conducting another meta-analysis on this topic would be the best use of available resources.

3.2 STUDY II

Study II is titled *Sex Differences in Episodic Memory Variance*. It was published in *Frontiers in Psychology* in 2020 (Asperholm et al., 2020).

3.2.1 INTRODUCTION

As already briefly outlined in section 1.9, the goal of study II was to investigate variance differences between men and women in episodic memory performance, mainly with regard to the type of material to be remembered. Here, it was hypothesized that in line with previous research on sex differences in both cognition and more general traits (see section 1.7), men would be more variable on both the full dataset as well as for all material categories.

3.2.2 METHOD

For this study, several five-level meta-analyses (further described in section 2.2 and 4.3) were conducted on a subset of the original dataset (the full dataset is described in section 2.1, and the subset is described further down), using the dependent variables Hedges' *g/lnVR* and the weighted variance of Hedges' *g/lnVR* (further described in section 2.2.2).

For these analyses, only data points where some kind of variance metrics were presented separately for men and women could be used (for example, if only pooled standard deviation was available for a data point, that data point had to be discarded). This resulted in a final dataset where 535 out of the 617 studies in the full dataset were included, totaling 962,946 participants.

3.2.3 RESULTS AND DISCUSSION

First, we performed a replication of an analysis in Study I (see Figure 6) with Hedges' *g* as the dependent variable and *type of material*

Type of material	lnVR	95% CI	VR	k	p	I^2
Verbal	0.018	[0.001,0.036]	1.02	333	0.04	76%
Images	0.009	[-0.014,0.033]	1.01	174	0.44	88%
Movies	-0.023	[-0.106,0.060]	0.98	19	0.59	19%
Locations	0.036	[0.001,0.070]	1.04	56	0.04	95%
Routes	-0.173	[-0.261,-0.084]	0.84	15	0.00	79%
Faces	0.039	[-0.006,0.084]	1.04	47	0.09	64%
Sensory	0.092	[-0.035,0.220]	1.10	7	0.16	45%
Remaining	0.029	[-0.014,0.071]	1.03	60	0.19	70%
Total	0.017	[0.000,0.030]	1.02	535	0.03	84%

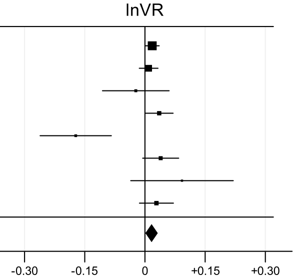


Figure 7: Figure 2 from Study II showing how the variance sex differences in episodic memory performance varies with the type of material that one is supposed to remember (Omnibus p<.01; I^2=83%). Each row shows whether the effect size for that level of the moderator is reliably different from 0. The last row show the result from a meta-analysis without any moderators, performed on the full dataset. Positive values mean that women performed better than men. Explanation of headings: *lnVR* = *lnVR*; *95% CI* = the 95% confidence interval; *k* = number of studies; *p* = the *p*-value; I^2 = statistic denoting the percentage of variation across studies that is due to heterogeneity rather than due to chance.

to be remembered as the moderator. This was carried out in order to make sure that the original results from Study I still held for the somewhat reduced dataset. Results (see Table 1 in Study II) showed that the categories *movies* and *remaining* went from being significant to non-significant.

Next, the same analysis as above but with *lnVR* as the dependent variable was carried out. The results (see Figure 7) showed that men were more variable for *verbal* and *locations*, while women were more variable for *routes*. Men were also more variable for the *total* dataset. What is interesting to note here is that when comparing the results of the *type of material* moderator analyses using Hedges' *g* as the dependent variable to the same analyses using *lnVR* as the dependent variable, it can be seen that they mirror each other for all level pairs that are significantly different from zero in both analyses. That is, whenever women performed better than men, men also had larger variance than women and vice versa (I will come back to this shortly).

Further, we performed a number of meta-regressions with either *age* or *publication year* (defined in section 2.1.3) as moderators on each *type of material* subset as well as on the *total* dataset. Results (see Table 2 and Table 3 in Study II) showed that *publication year* had no effect on the outcome. Further, for *age*, two small, negative linear relationships could be found, here indicating that the men's larger variances compared to women became smaller across age.

Next, a number of meta-analyses were carried out in order to search for possible bias in the dataset, using the following four moderators (defined in section 2.1.3): *database search* (whether data was taken from the first or second database search), *data source* (whether data points were taken from publications or received from authors), *study objective* (whether sex differences explicitly were investigated), and *sampling of subjects* (whether the sample was convenience-based or population-based). Similarly to Study I, only *sampling of subjects* showed a significant omnibus test (see Table 4 of Study II), meaning that the sex difference in variance was more positive (that is, towards men having larger variance) in convenience-based samples compared with population-based samples. Otherwise, there were no indications of bias in the dataset. As already mentioned in the results section of Study I, all data from the open databases were categorized as population-based, which means that possible deviations in these have a large probability of showing up for the *sampling of subjects* variable as well.

Returning to what was discussed earlier, a somewhat interesting and consistent, mirrored pattern appeared when comparing the two *type of material* moderator analyses performed with Hedges' *g* and *lnVR* as dependent variables respectively: That is, whenever men performed better for a material category, women also had larger variance than men and vice versa. This pattern was further investigated in a number of exploratory meta-regressions performed separately for all the *type of material* subsets as well as the *total* dataset, where *lnVR* was set as the dependent variable and Hedges' *g* as a moderator. Results (see Table 5 in Study II) showed significant, positive relationships for the *verbal*, *images*, *locations*, *routes*, *faces*, and *total* sets, and a significant, negative relationship for the *sensory* subset. That is, there were positive relationships found between the mean differences and variance differences for all categories where the mirrored pattern described above could be identified. A positive relationship here indicates that as the performance goes toward favoring women on a given task, the variance of men becomes larger and larger compared to the women's and vice versa (for an illustration of this relationship for the full dataset, see Figure 8).

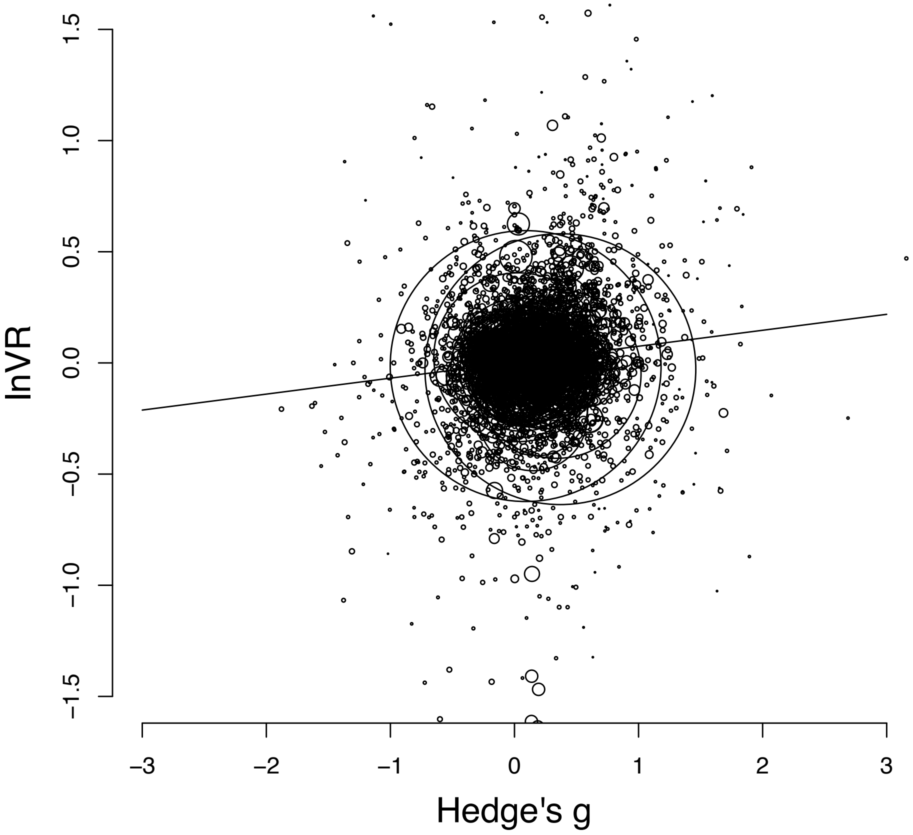


Figure 8: Figure 3 from Study II showing the relationship between *lnVR* (where positive values mean that males have more variance than females) and Hedges' *g* (where positive values mean that females perform better than males) for the full dataset. The diameter of each data point is equal to the inverse of its squared weighted variance (see section 2.2 for an explanation of this concept). The drawn line shows the best-fitting regression line from the significant meta-regression reported in Table 5 of Study II.

One way of explaining this pattern could be that it comes about because of ceiling effects. That is, when one sex performs much better than the other, they may also hit a ceiling, either an actual limit to their performance or an upper bound resulting from how the test is designed; an example of the former being if they were asked to recall fewer items than what they actually were capable of remembering. Since the performance distribution (regardless of the underlying reason for this ceiling effect) then would be restricted for the top achievers, the variance would also become narrower. This, in turn, would mean that the sex who would be less affected by this ceiling effect also would have larger variance than the other.

If a ceiling effect like the one described above would come about because of an actual limit in the cognitive performance, this would also reflect an actual difference in the variance between men and women. Such a limit — or diminishing returns effect if you will — is something that is bound to happen at some point if the performance of one sex keeps increasing in relation to the other, the question is only when. However, if it would come about because of deficiencies in how the outcome was measured, the obtained results would only be artifacts, not indicating a real phenomenon in the underlying performance. With the data we have, finding out which of these factors contribute to the outcome to what degree (if at all) is not possible. However, it *is* possible to point to other large studies where the mean and variance differences go in the same direction (Hedges & Nowell, 1995; Lakin, 2013; Machin & Pekkarinen, 2008; Nowell & Hedges, 1998), indicating that *even if* a ceiling effect could be a factor in some cases, it cannot fully explain why men seem to be more variable than women in many areas.

Finally, an illustration of what the overall distributions for men and women would look like, given that the estimations from the two *type of material* moderator analyses presented above were accurate, is shown in Figure 9. Here, while it can be argued that the variance differences are quite small, it can also be seen that the distributions of men and women in the extremes are not negligible. For example, for the *verbal* category, there are 1.76 women for each man in the top 5%.

In summary, the small effect sizes that were shown combined with some question marks surrounding possible ceiling effects in the underlying data makes this study informative but also warrants further research. More specifically, it would probably be useful to investigate

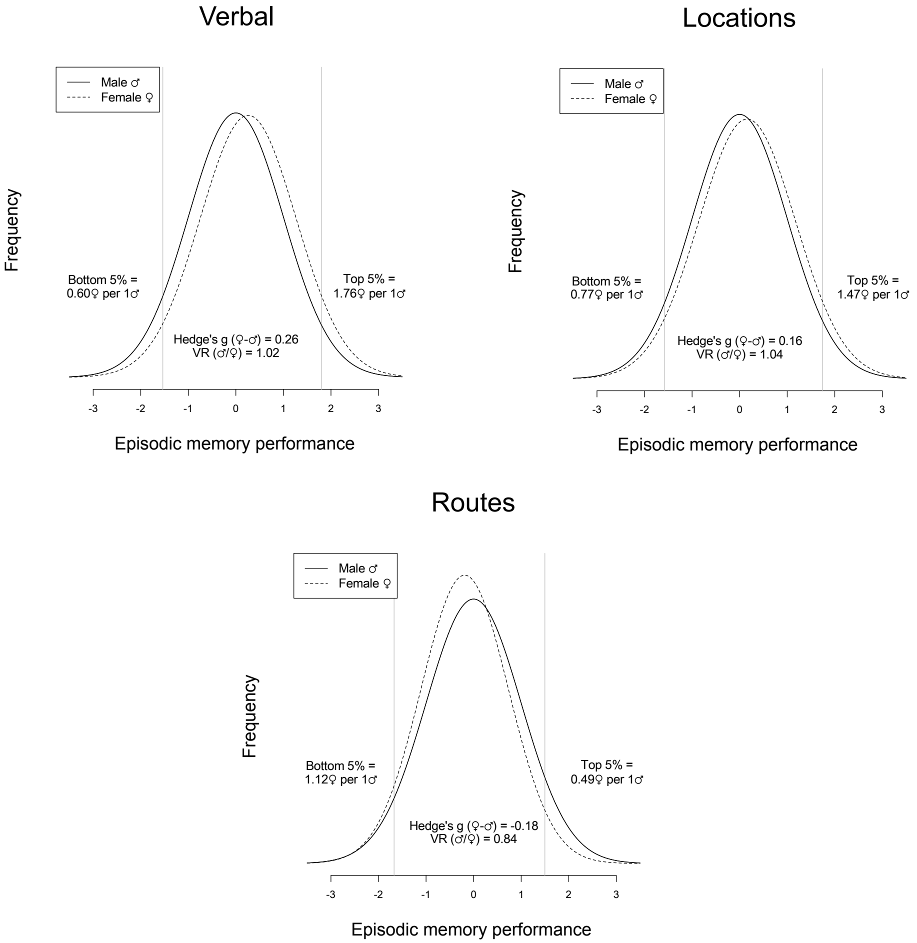


Figure 9: Figure 4 from Study II that shows assumed distributions of male and female performance for the *verbal*, *locations*, and *routes* subsets. These distributions are computed based on the estimates of Hedges' *g* and *lnVR* in the respective analyses where these were assessed with *type of material* as a moderator (see Table 1 of Study II and Figure 7 of this thesis). Assumed distributions are only shown for *type of material* categories where both of these estimates significantly differed from zero.

a large dataset containing the raw participant data, rather than just descriptive summary statistics, meaning that it would be possible to say more about potential problems in it. Even so, the study does provide some circumstantial evidence that what has already been shown for other domains, namely that males tend to vary more as a group than females, also might be the case for episodic memory.

3.3 STUDY III

Study III is titled *The Magnitude of Sex Differences in Verbal Episodic Memory Increases with Increased Gender Equality: Data from 54 Countries Across 40 Years*. It was published in *PLOS ONE* in 2019 (Asperholm, Nagar, et al., 2019).

3.3.1 INTRODUCTION

As already briefly outlined in section 1.9, the goal of study III was to investigate if men and women are affected differently by societal progression when it comes to mean differences in episodic memory performance. More specifically, we first wanted to explore whether country moderated the mean sex difference effect in episodic memory performance. If this would be the case, we then wanted to go on and investigate possible predictors that could make sense of this pattern. The hypothesis was that country would moderate the mean sex difference effect and that social progress would be positively related with changes in sex differences to women's adgantage. Further, we hypothesized, in line with earlier research (see section 1.7.3), that a measure of gender equality would be the strongest social progress predictor for this relationship among the measures investigated.

3.3.2 METHOD

For this study, we used the dataset described in section 2.1. However, only studies that had been performed within a single, known country were included, resulting in a final dataset of 612 studies, originating from 54 different countries, published between 1973 and 2013, and involving a total of 587,691 participants. Further, the dataset was divided into *verbal* and *non-verbal* material. Here, *verbal* material was defined the same way as in the *type of material* variable (see section 2.1.3), with the addition that nameable images (see section 3.1.2) also were included. This decision was made in light of the results of Study I, where women did not only perform better than men for nameable images, but where there also was a significant difference between nameable images and non-nameable images. In addition to their theoretical closeness, this made nameable images similar to the verbal material category from an empirical standpoint as well; when remembering words, they can often be remembered as images, and vice versa.

Further, three indicators of social progress were constructed, where each unique country and year combination received a value based on what the numbers were at that specific point in time for that specific country. Data for these measures were taken from *United Nations Development Programme* (UNDP) and *The World Bank* (WB). The three indicators were defined as follows:

> **Gender Equality.** A composite measure that was made up of (1) the UNDP indicator *Average Years of Schooling Attained (Female population, 25 years and over)* subtracted by *Average Years of Schooling Attained (Male population, 25 years and over)* and (2) the WB indicator *Female to male ratio on labor force participation (for ages between 15 and 64)*.
>
> **Population education and employment.** A composite measure that was made up of (1) the UNDP indicator *Average Years of Schooling Attained (Total population, 25 years and over)* and (2) the WB indicator *Labor force participation rate (% of total population ages 15-64)*.
>
> **GDP per capita.** Gross domestic product per capita is a measure of economic activity in relation to the size of the population. This measure, given in current US dollars, was extracted from the WB database and subsequently transformed using the natural logarithm to achieve a normal distribution of the variable.

Indicators were not always available for every year. Whenever an indicator was missing, it was extrapolated from indicator values surrounding it. Further, each indicator was z-transformed using the full distribution of all possible year and country combinations for the included countries between 1972 to 2013. Also, correlation analyses for the full dataset revealed positive associations between the three indicators (*Gender Equality* vs. *Population Education and Employment*: $r=.70$, $p<.001$; *Gender Equality* vs. *GDP per Capita*: $r=.50$, $p<.001$; *Popula-*

tion Education and Employment vs. *GDP per Capita*: r=.56, p<.001).

When it comes to indicators measuring different types of social progress and living conditions constructs, such as the ones presented above, there are a lot of choices to be made. For example, one of the most popular measures with regard to gender equality is the *Global Gender Gap Index* (GGGI) (World Economic Forum, 2020). This is a composite measure consisting of a number of variables that together are meant to gauge women's opportunities/participation when it comes to economy, politics, health and education. However, it can be questioned whether this scale actually captures gender equality, and other measures have been proposed (Stoet & Geary, 2019). One critique here is, for example, that within the scope of the GGGI, it is theoretically impossible for women to be better off than men because underlying measures are capped once they reach perfect symetry between the sexes. It also does not focus on areas where men usually are worse off such as harsher punishments when it comes to crime, higher suicide rates, and more occupasional accidents (Stoet & Geary, 2019).[19]

For the present study, we needed to find indicators that not only captured the underlying concepts that we were interested in, but also that suited our data fairly well when it comes to available data for the specific countries and year that we happened to need. As such, we chose to go with the relatively simple measures specified above.

3.3.3 RESULTS AND DISCUSSION

First, in order to evaluate whether country moderated the episodic memory effect size, two five-level meta-analyses with country as a moderator were performed, one on the *verbal* subset and one on the *non-verbal* dataset. Results (see Figure 10) showed that for the *verbal* subset, country was significantly modifying the effect sizes. However, for the *non-verbal* subset, it did not. Therefore, going further, only the *verbal* subset was considered.

Next, we fitted a number of six-level meta-analyses (where country was added as the sixth level on top of the hierarchical structure described

[19] However, it can, of course, also be questioned to what extent actual outcomes measures oppurtunities. Does low policital participation from women automatically mean that they are held back? Does higher suicide rates for men mean that society somehow has done something wrong?

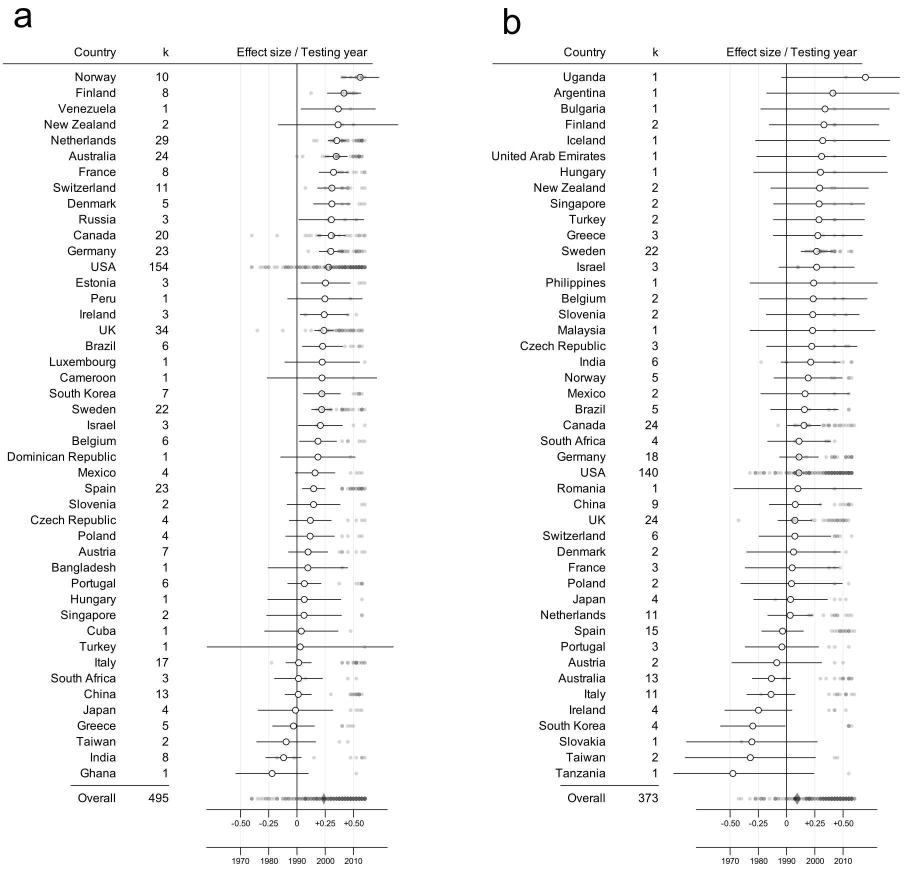


Figure 10: Figure 1 from Study III (taken from the correction of the original article; The PLOS ONE Staff, 2019) showing the results of two moderator analyses (investigating the *verbal* and *non-verbal* sets of the data) with country as a moderator and Cohen's *d* of episodic memory performance as the dependent variable. Number of studies is indicated by the *k* column. Positive values indicate that women perform better than men.

in section 2.1.2): The first thing we did was to fit three simple models with each of the social progress indicators as moderators and then a combined model with all three of them included simultaneously. Here, Cohen's *d* of episodic memory performance along with the weighted variance of Cohen's *d* were the dependent variables, and each effect size was paired with the right social progress indicator value according to when[20] and where the study was carried out. The results (see Table 1) of the three simple regressions all showed significant, positive relationships between the dependent variable and the indicator (see Figure 11 for illustrations of these effects). That is, they all went in the expected direction where an increase in women's relative episodic memory performance compared to men's was related to increased social progress. However, for the combined model, comprising all three social progress indicators, a significant, positive relationship could only be found for *Population Education and Employment*. Here, one *z*-point change for this indicator was associated with an increase of 0.08 in Cohen's *d* for verbal episodic memory performance.

In addition, we performed the same meta-regressions as above for *verbal* episodic memory tasks, both *without* the large databases that were the sole focus for Weber et al. (2014) and Bonsang et al. (2017) as well as with *only* these databases. This did not change the outcome in a major way (see S2 Table of Study III). For the analysis without the databases, *Population Education and Employment* went from being significant to marginally significant for the combined model. For the analyses with only the databases, *GDP per capita* went from not being significant to actually being significant.

The upshot of the results of this study is that while all social progress indicators tested for could be shown to be positively related to women's relative increase in verbal episodic memory performance compared to men's, it was the overall education and employment level rather than the gender equality level that best predicted this. Given the reasoning laid out in section 1.8.3 — where it is argued that increases in living standards should benefit women more since men might experience a form of diminishing returns effect from already having access to more

[20] Here, we used the publishing year of the article in question minus one to better reflect when the experiment in the study probably was carried out. Just subtracting by one was probably too conservative though, and in retrospect I believe that we could have increased this number. I also believe that this would not affect the overall results.

	SM1	SM2	SM3	CM
Intercept				
Estimate	0.16***	0.13***	0.18***	0.14***
Standard error	0.02	0.02	0.02	0.02
Gender Equality				
B estimate	0.10***			0.04
B standard error	0.02			0.03
Population E&E				
B estimate		0.13***		0.08*
B standard error		0.02		0.04
GDP per Capita				
B estimate			0.60***	0.18
B standard error			0.16	0.20
Observations	2681	2681	2681	2681
R²	0.12	0.15	0.11	0.16

Table 1: Adapted version of Table 1 in Study III showing summary statistics of the four meta-regression analyses (SM = Simple model; CM = Combined model) for *verbal* episodic memory performance. Explanation of symbols: $* \rightarrow$ *p*<.05; $** \rightarrow$ *p*<.01; $*** \rightarrow$ *p*<.001. Notice that there is a typo in the original article that has been corrected here: The *Gender Equality B* estimate for SM1 should be *p*<.001 rather than *p*>.05.

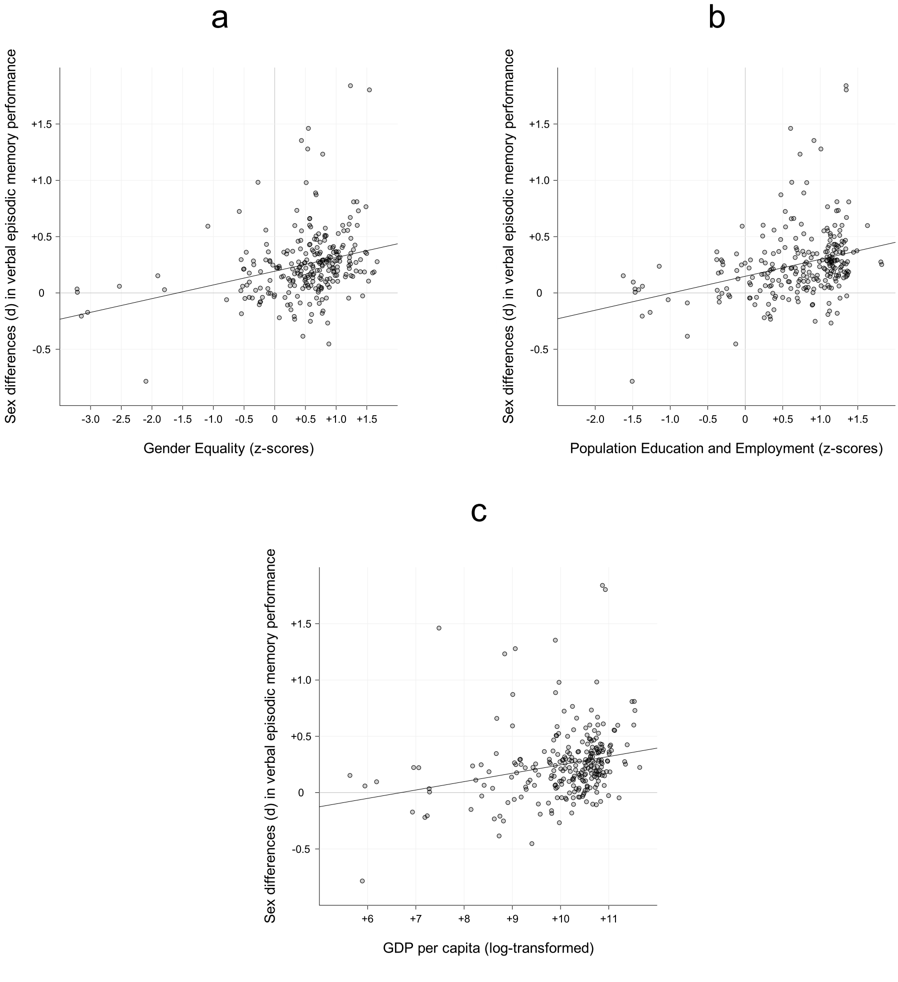


Figure 11: Figure 1 from Study III (adapted from the correction of the original article; The PLOS ONE Staff, 2019) showing the best-fit regressions of the three separate meta-regressions with sex difference in *verbal* episodic memory performance as the dependent variable and the social progress indicators (a) *Gender Equality*, (b) *Population Education and Employment*, and (c) *GDP per Capita* as continuous moderators.

social goods — it would make sense that increasing the general level of education and employment, independent of sex, could benefit women more than men as long as they were not left out completely.

It is also possible to construct a hypothetical scenario showing how gender equality could be less important: Imagine two countries, Oceania and Eurasia. In Oceania, the general standard of living has increased dramatically while gender equality has seen less progression. In Eurasia, the general standard of living has been quite unimpressive while gender equality has seen a striking improvement. In this scenario, women in Oceania would probably experience larger increases in cognitive performance compared to men than what women in Eurasia would. If our dataset were to share characteristics with this example, it could be argued that this at least in part could explain the pattern seen.

In summary, with this study, we have further strengthened what others already have suggested, namely that social progress is related to relative improvements in sex differences to women's advantage, both in cognition in general and for episodic memory. However, the fact that it was the overall education and employment level rather than the gender equality for these measures that was the best predictor for the outcome suggests that more in-depth studies, probably applying some other method than what has been used so far, are needed in order to better explain what is going on.

4 DISCUSSION

For this thesis, I have through three different studies investigated sex differences in episodic memory from several, disparate perspectives. Rather than setting up experiments and gathering participant data ourselves, we synthesized and combined already produced data in order to reach new and powerful conclusions.

In Study I, we performed a meta-analysis on mean sex differences in episodic memory, using a dataset of 617 studies that we compiled for this purpose, consisting of both published and unpublished data, as well as several open databases. Here, we could first and foremost, as expected, see a strong indication that just like with other cognitive tasks, the difference in performance between men and women depends on what type of material they are being exposed to. More specifically, men tend to perform better on more spatial tasks while women tend to perform better on more verbal tasks, and this pattern also seems to be in place for episodic memory. Women also performed better than men for tasks having to with remembering faces as well as assignemtns pertaining to smell, touch, and different shades of colors. These type of tasks cannot, at least not on the face of it, be categorized as either verbal or spatial and would therefore suggest a broader advantade for females when it comes to episodic memory.

In Study II, we performed a number of analyses on the same dataset gathered in Study I, searching for variance differences between men and women. Here, we could, in line with our expectations as well as previous research, find that men overall were slightly more variable than women when it comes to episodic memory. However, we also found some exploratory results that implied that those findings might come about because of underlying methodological problems in the original research, more specifically possible ceiling effects in a subset of the tasks. Therefore, more research to investigate these potential problems is warranted.

In Study III, we performed a number of analyses on the same dataset gathered in Study I, investigating whether sex differences were related to social progress. Here, we could first see that which country a study was conducted in also affected the sex difference in verbal episodic memory. Investigating this further, we concluded that sex differences in verbal episodic memory tracked several different indicators of social progress tied to the year and country of each individual study. However, when pitted against each other, it was only the education and employment level that could be shown to have an effect and not gender equality, which we expected would be the most important indicator.

It should be mentioned that while the sex differences and effects that are demonstrated within the scope of these three studies often are quite small, one should keep in mind that they also represent basic cognitive skills that form the basis of a plethora of higher-order cognitive functions. In that respect, it is not unreasonable to suspect that these small differences will be amplified and result in much more obvious differences for more complex behaviors. For example, when reading a text, episodic memory is involved for multiple tasks that you have to carry out, for example encoding what you currently are reading, recall-ing what you read previously, as well as remembering newly defined terms and concepts. This means that performing just slightly better in episodic memory also will impact all of these tasks which in all individually increases the overall reading performance. When combining this with repeated exposures and training over a long time, even small effects for basic skills should translate to larger effects for more composite skills.

I will now turn to discuss some general aspects of the studies included in this thesis more in-depth.

4.1 QUALITY VS. QUANTITY

When collecting a large amount of data in order to get a full picture of the research that has been done within a field (as we did for the studies in this thesis), one often has to make a sort of quality/quantity decision. That is, on the one extreme, for our data collection we could have chosen a single standardized test of episodic memory and then only gathered data for that specific measure. This has the advantage that you can be rather confident that you are not comparing apples to oranges. Instead, you will investigate a very specific measure that probably also corresponds to a rather narrow construct, meaning that you might inspect just a sliver of the construct that you really are interested in. This means, in turn, that your results, in theory, should be highly reliable (as long as you find enough articles, which might be a problem if you are using a very narrow measure) while the validity concerning the more encompassing construct might be quite weak.

On the other extreme, which is closer to what we actually did, you can chose to collect any measure that in some way relates to the construct that you are interested in, meaning that you will cover many different aspects of it, something that in theory should strengthen your validity. In addition, since a bigger set of possible articles to draw from means that you can get a larger sample, the reliability should also increase. However, even if you have a larger sample, the fact that you also have more tasks that might relate to the underlying construct to different degrees means that even a balanced dataset (that is, where all type of tasks are evenly represented) can lead to low validity. To combat this, it would be possible to, for example, weight tasks differently depending on how well they relate to the underlying construct. However, this has the risk of becoming too arbitrary if one does not have a clear and well structured idea of how to carry it out. Here, it might therefore become necessary to divide the data into different subsets for further analyses, something that we did ourselves with out *type of material* to be remembered division (see section 2.1.3).

It should also be mentioned that there, of course, are other ways to affect either the reliability or validity, regardless of this quality/quantity aspect. For example, one could increase the reliability by applying some sort of quality judgments of the studies that in turn would weight articles differently. However, this would be very time consuming and might be less effective than simply collecting more articles, if that is a choice that can be made instead. The reasoning here could be that 100 rather low-quality studies still might give a better final estimate than 10 high-quality studies, given that the error variance for the former is non-systematic in nature.

Depending on factors such as time and amount of data available, and how clear of an idea that can be formed of the underlying construct, one has to make a number of informed and pragmatic decisions regarding the alternatives listed above. As described in section 1.4, episodic memory is a rather broad concept, referring to a set of multi-factorial/multi-modal memory processes, and it is therefore hard to encircle it using a narrow set of measures. Furthermore, to the extent that it is possible to construct relatively exhaustive tests of episodic memory, taking into account different modalities and aspects of it, such tests are fairly uncommon in comparison to more simple memory tasks that to some extent can be seen as proxies to episodic memory, approaching it from one of the many angles. So when choosing a rather liberal set of measures to represent episodic memory, there also is a very large set of studies conducted that can be included. These factors, among others (for example more practical concerns regarding how much time that could be allocated to each phase of the data collection), contributed towards us choosing a rather broad/permissive/liberal set of measures and deciding against judging each study based on some form of quality scale, opting to gather more studies instead. This means that the main strength of the dataset (with regards to the perspective discussed in this section) is its sheer size rather than its extremely precise nature.

4.2 BIAS AND QUESTIONABLE RESEARCH PRACTICES

Rather recently, psychology as a field has dealt with what some like to call a replication crisis, that is, a situation where it has been shown that a lot of the findings probably will not replicate.[21] While this discussion has been going on for a long time (see, for example, Francis, 2012; Ioannidis, 2005b), with the underlying mechanisms for why this might come about having been known since the birth of modern statistics, one of the articles that really have helped to highlight this issue came out in 2015 (Open Science Collaboration, 2015). Here, 100 research teams conducted exact replications of 100 different findings published in top-ranked psychology journals. Replication success was tested in several different ways. For example, out of the 97 studies that originally had demonstrated significant results, only 36 of the replications showed significant results as well. Also, the average effect size for the replications was about half the size of the average effect size for the original studies. One way of explaining why these replications failed to such a large degree is that the researchers in question simply were not good enough at conducting the replications, thereby failing at it. However, speaking against this interpretation, most experiments were conducted in collaboration with the original authors. Further, it could also be shown that variables like the magnitude of the original effect size were better predictors of whether the research would replicate than, for example, the expertise level of the research team.

If it actually is the case that many results within psychology are false, several different factors can be put forth to explain why that is. In this regard, one sometimes talks about different *questionable research practices* (QRP). These include things like, for example, change or add to the statistical analyses that are being performed until positive findings show up, selectively reporting only dependent variables and analyses that showed positive results, changing the hypotheses after inspecting the data, choosing how to deal with outliers based on how they affect the outcome, and continuously running statistical tests and stopping the data collection once results are satisfactory. Researchers can also choose not to publish negative results, or even be denied by journals if trying, which can erroneously skew overall findings of a field.[22] These

[21] However, this discussion is not unique to psychology (see, for example, Baker and Penny, 2016; Begley and Ellis, 2012; Camerer et al., 2016; Ioannidis, 2005a).

[22] As an analogy, if a fisherman kept all the tunas that he caught but threw back all

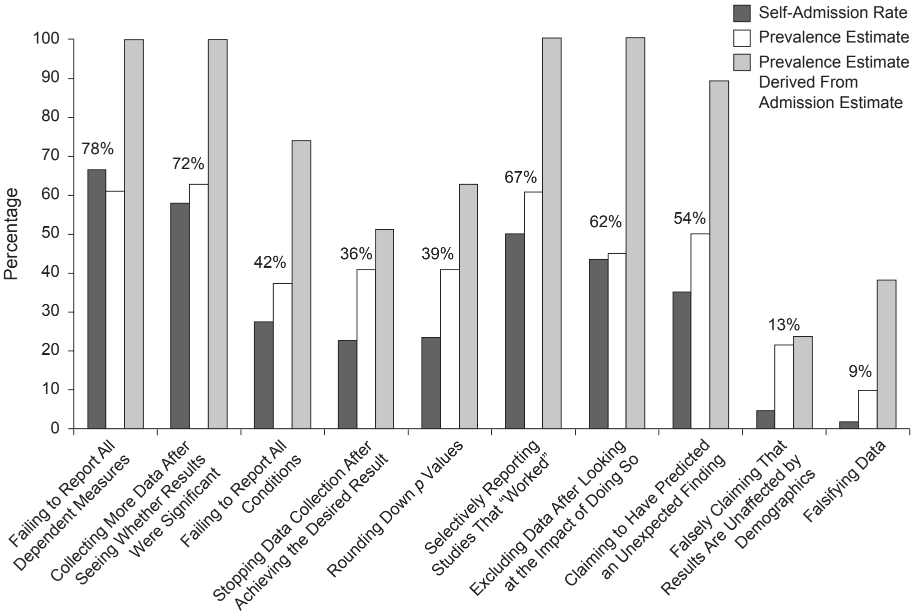


Figure 12: Figure 1 from John et al. (2012) where the black bars indicate how large proportions of researchers (from a sample of 2,155 academic psychologists in the U.S.) admitted to having engaged in the QRP in question. White bars indicate their estimation of the proportion of other academic psychologists that had engaged in it.

QRP can range on a spectrum from being relatively mild (for example overselling the findings) to outright cheating (for example inventing data; Neuroskeptic, 2012). In a survey carried out by John, Loewenstein, and Prelec (2012), it could be shown that many QRP are relatively common among researchers (see Figure 12), with for example almost 70% (the largest self-admission rate found) admitting to not always report all dependent measures that they collected.

No matter the underlying reasons, it can be concluded that if the available research in a certain field were to deviate too much from the actual truth, this would cause major problems when it comes to synthesizing research, such as the studies in this thesis. One should therefore ask to what extent different types of biases and QRP might influence the results of this thesis. Thankfully, there are a number of circumstances

the pikes, examining what type of fish he brought back to shore would not be a good strategy for drawing a conclusion of the ocean fauna. Worse yet, if some of the fishes just were hallucinations (which could be argued to be the case when QRP have been heavily used), this certainly would not help.

that makes it possible to say something about this:

> **Several successive outcomes.** Within many research fields, the outcome of a study can often only be positive in one specific way while still being consistent with the underlying theory. That is, while it certainly is theoretically possible that cognitive behavioral therapy could affect phobias negatively, such a finding would probably be seen as a failure in finding something real. As such, QRP would probably not tend to lead to this outcome, and studies showing this would presumably not be published.
>
> However, when approaching the research question if there are sex differences for a certain skill or trait, probably all three major conclusions — that women perform better, that men perform better, and that there are no sex differences — could potentially be counted as positive findings by most researchers. As such, even if there were no sex differences, and even if QRP were in full effect, one would not expect either sex to ostensibly perform better. Rather, studies showing either that men did better or showing that women did better should be available to about the same extent. Bias inducing mechanisms are therefore less probable when finding overall evidence that a certain sex performs better than what would be the case in many other fields. Or at least, the bias inducing mechanisms would have to be of a more complicated nature, which is less probable.
>
> **Studies with other research questions.** As already mentioned in section 2.1.3 and presented in the result sections of Study I and Study II, not all studies included in the dataset had the explicit goal of investigating sex differences. Rather, in a majority (about 58%) something else was investigated, suggesting that QRP should not be present for these, at least not in any obvious way. This means that from the start, about half of the dataset probably is not even susceptible to this type of bias, and as shown in the bias analyses for Study I and Study II (see the results sections of these studies), whether sex differences were investigated could not be shown to be an influential moderator.
>
> **Unpublished findings.** As already mentioned in section 2.1.1, about 40% of the data was sent to us from authors after having queried them via email about this possibility. That is, this was previously unpublished results that the authors themselves had

not summarized and presented in the article. If data points from studies where other research questions have been investigated are somewhat protected from QRP, data points from studies where the data was not even reported (and presumably, in a lot of cases not even considered) should be even more protected. Obviously, there is an overlap with the previous concept here, meaning that most studies received from authors also fall into the category of studies not investigating sex differences. In addition, both for Study I and Study II (as shown in the bias analyses in their respective result sections), whether the data was taken directly from articles or received from authors could not be shown to affect the effect size estimates.

Taken together, the reasons listed above should at least alleviate concerns one might have about the underlying data being biased from QRP in some way.

4.3 DEPENDENT DATA PROBLEM

As already mentioned, in section 2.1.2, the dataset that we used had a form of hierarchical structure, with dependent data points. However, in a standard meta-analysis, one of the many assumptions concerning the underlying data is that each data point is independent from all the rest. This means that in the total set of effect sizes, $\{e_1, e_2 \ldots e_n\} \in E$, there should be no clusters of effect sizes, $\{c_1, c_2 \ldots c_n\} \in C \subseteq E$, where knowing the value of at least one of the cluster effect sizes, c_x, will enable you to better predict any of the remaining effect sizes in the cluster, $C \setminus c_x$, than any of the effect sizes outside of the cluster, $E \setminus C$. A simple example where this assumption would be violated, even in a dataset where each study only contributes with one data point, is if the same sample of participants has been used in multiple studies. In this case, knowing the effect size from one study where this sample was used would help to better predict other effect sizes where this sample also was used in comparison to effect sizes where it was not used.

This whole dependence issue is a problem because the underlying model in a standard meta-analysis does not take it into consideration. For example, imagine that you had three data points, each derived from different studies, but with equal number of participants, where the standardized differences (that is, Cohen's *d*) are -1, 0, and +1 respectively. Given these numbers, you would have to estimate the true mean difference for the population as 0. But what if you find out that the first and second data points actually come from the same sample? In your original prediction, you gave that sample an unfair weight since it contributes toward two effect sizes and the other only one. To make this point even more clear, imagine that you had 999 effect sizes from one sample and 1 effect size from another. Giving the last effect size the same weighting as any other would render it more or less meaningless in the final estimate even though it would contribute with half of the total amount of participants.

Now, it is important to realize that in practice, clusters like these will always exist in the data, even if they are not as obvious as in the examples above. For example, effect sizes could originate from research performed in the same country, from researchers adhering to the same methodological school, or from experiments conducted in rooms with the same color on the walls. In the end, what constitutes dependency comes down to how much error one is willing to accept. That is, if the color of the room is perceived to only affect the result in a minuscule way, one can ignore this aspect. Most commonly, however, researchers tend to assume that data points are independent from each other as long as each data point is taken from a unique study.

As already mentioned, there is a clear and very influential dependency structure in our dataset where each study can contribute with several different samples, where each sample can contribute with several different encoding sessions, and where each encoding session can contribute with several different measures. Fortunately, there are several ways to deal with this problem (Hedges, Tipton, & Johnson, 2010; Scammacca, Roberts, & Stuebing, 2014):

> **Ignoring it.** One solution to the problem is to simply ignore it and go ahead with the analyses anyway, something that might not be that big of a deal if only a few data points are clustered. In certain situations, this path can even give conservative results that can be used to get a rough overview of the underlying patterns (Hedges et al., 2010). However, this method is of course not recommended when a high level of accuracy and power is needed.
>
> **Data reduction.** The simplest and most straightforward way to actually handle dependent data points (in contrast to just ignoring the problem) is to perform some kind of data reduc-

tion/combination procedure to get a dataset without any dependencies that is suitable for the standard meta-analytic method. This could be done either (1) by choosing the most relevant data point from each study or (2) by combining data points within each study in order to end up with non-dependent measures (here, one would sequentially have to combine data points on each hierarchical level, gradually merge combined data points from lower levels until a final measure is achieved). Regarding the first option, this would, in the case of multiple outcome measures, require one to be able to understand which one of these that is the most relevant, and in the case of several samples, one would simply have to keep only one. Discarding data in this way is a problem since this lowers the precision of the final measure. That is, imagine that you measured something 100 times and then were forced to just pick a single data point to keep. Obviously, as long as all type of measurements are relevant, having more of them would enable you to be more sure about the underlying distribution they were sampled from.

Another problem here is that in order to derive a final variance measure for the combination procedure, one also has to specify an assumed correlation between the data points. If this correlation is set to 1, it does not matter how many times you measure something; the final measure will not be more precise than if you only measured it once. On the other hand, if this correlation is set to 0, gathering more data points will always increase the precision of the final measure. Correlations between data points can often be hard to know (or even hard to just estimate), and they might very well differ between different tasks, samples, and studies. This means that when not having access to the original raw data of each study (which one very seldom do), the best one can do is to err on the side of caution and choose a correlation that almost surely is too conservative.

Finally, when using one of these methods, in order to carry out moderator analyses with not more than two levels, one would first have to derive two separate measures for each level within each cluster (that still would be dependent in relation to each other) and then subtract one from the other. This would result in a difference score for each cluster, that then could be used in a standard moderator meta-analysis. However, performing moderator analyses with moderators with even more levels is not possible when using this method.

Specifying the covariance structure. If the covariance structure between dependent data points, discussed above, is known or if a reasonable estimation of it can be made, it is possible to to use it within a multivariate model directly (see, for example, Hripsime and Raudenbush, 1996). Here, if this covariance structure is accurate enough, this method is the also the most accurate one to use. However, as already touched on above, unless one has access to the raw data, making the necessary estimations are both very hard and time consuming. It is, however, also possible to conduct sensitivity analyses, using different covariance structures to see how the result may vary when utilizing different estimates.

Leaving out the covariance structure. As discussed several times above, finding out or estimating the covariance structure between dependent data points can be an extremely time-consuming task, and in many cases it can be almost impossible to derive good enough estimates. In these cases, there are two methods that can be used: *Robust variance estimation* (Hedges et al., 2010) and *n-level meta-analysis*[23] (Konstantopoulos, 2011). Both these models can, albeit differently, handle not having access to what degree dependent data points covary, and also handle hierarchically clustered data.

Hoare (1980, p. 81) concluded that "there are two ways of constructing a software design: One way is to make it so simple that there are *obviously* no deficiencies, and the other way is to make it so complicated that there are no *obvious* deficiencies." The same thing goes for statistics where you often have to choose between methods that are simple/easy to understand and methods that are more complicated/hard to understand. As such, going with the simple alternative might be less accurate but more transparent and easier to assess by a reader. On the other hand, going with the more complicated option might be more accurate (if one applies it correctly, which might be harder to do) but

[23] The term *n* here stands for the number of levels that the model has, where the most simple meta-analytical model has two: One modeling the between-studies variance and one modeling the within-studies variance (or error variance). As such, when having clusters of studies from, for example, different research labs, this would warrant a three-level meta-analysis (with research labs being the third one).

less transparent and harder to assess for the recipient.

For the data that constituted the basis of this thesis, going with any of the first two methods (ignoring the problem or perform some kind of data reduction) would, to paraphrase Hoare, be *obviously* wrong; using the data reduction method would still mean that data points were dependent for moderator analyses with more than two levels.[24] On the other hand, the third alternative (specifying the covariance structures) would require information that simply was not available, neither could it be estimated in a reasonable way. Remember that the large heterogeneity of the different types of tasks would make it necessary to come up with about as many dependency structure estimations as there are unique tasks and ways of measuring the outcomes of those very tasks.

Thus, we are left with the fourth, fairly complicated alternative (leaving out the covariance structure), which might be quite hard for the general reader to grasp. Here, I opted for n-level meta-analyses, somewhat because of the incidental circumstances that a reviewer of Study I suggested it and that I already had some experience using the *Metafor* package (Viechtbauer, 2010) for *R*, which has the capacity to carry out these type of analyses.

4.4 ETHICAL CONSIDERATIONS

Before ending this thesis, I should spend some time reflecting on possible ethical problems with the studies that we conducted when it comes to *how* it was being done (while section 1.3 more deals with aspects of *why* it should be done). In Study I, II, and III, all analyses are based on the same dataset, a dataset that is made up of aggregated numbers from already published articles and open databases. In some cases, we received data from authors, but the form of this data did not differ from what already was published (that is, aggregated numbers describing groups). Further, the open databases we used were, as the name suggests, more or less open already, available for just about anyone who can manage the hassle of going through a short registration process.

[24] Even if, for Study I, we used the data reduction method (more specifically, the combination procedure) to compute descriptive effect sizes for each study in Table S2 (not included in this thesis; see Asperholm, Högman, et al., 2019) and to generate the data points that made up the basis for the funnel plots for each material category subset in Figures S2 to S10. The exact method that was used for this combination procedure is described in the caption of Figure S1 (which *is* included in this thesis).

As such, the data and results are not any more traceable back to the original participants than they were in their original form in the different articles and open databases (if anything, it is harder). No ethical concerns with regard to the anonymity of participants should therefore be present for these studies.

However, while the tasks that were investigated in our dataset can be said to probably not cause any harm to the participants, it is of course impossible to guarantee this or that other unethical research practices were not utilized. Scientific journals have certain standards that the authors need to follow in order to get published, but these standards can still be circumvented since no rigorous control normally is being performed. As such, it *could* be the case that we, to a certain extent, base our findings on unethical studies. However, we have no indications that this would be the case, and even if it unbeknownst to us were the case, it is debatable to what extent this would further hurt the participants.

4.5 CONCLUDING THOUGHTS

As with all scientific endeavors, pretty much nothing is ever the final nail in the coffin for anything. Within the span of a Ph.D., you can often, at best, hope to nudge the knowledge within a field in one way or another. For this thesis, I hope (and feel) that I have contributed to this nudging process. The results presented might not have been extremely novel, but they have nevertheless hopefully helped us to get a more firm grip about the state of things when it comes to sex differences in episodic memory.

REFERENCES

Andreano, J. M., & Cahill, L. (2009). Sex influences on neurobiology of learning and memory. *Learning & Memory, 16(4)*, 248–266. doi:10.1101/lm.918309

Apostolou, M. (2013). Interfamily conflict, reproductive success, and the evolution of male homosexuality. *Review of General Psychology, 17*(3), 288–296. doi:10.1037/a0031521

Asperholm, M., Högman, N., Rafi, J., & Herlitz, A. (2019). What did you do yesterday? Sex differences in episodic memory. *Psychological Bulletin, 145*(8), 785–821. doi:10.1037/bul0000197

Asperholm, M., Nagar, S., Dekhtyar, S., & Herlitz, A. (2019). The magnitude of sex differences in verbal episodic memory increases with social progress: Data from 54 countries across 40 years. *PLOS ONE*, *14*(4), 1–11. doi:10.1371/journal.pone.0214945

Asperholm, M., van Leuven, L., & Herlitz, A. (2020). Sex differences in episodic memory variance. *Frontiers in Psychology*, *11*(613), 1–10. doi:10.3389/fpsyg.2020.00613

Astur, R. S., Ortiz, M. L., & Sutherland, R. J. (1998). A characterization of performance by men and women in a virtual Morris water task: A large and reliable sex difference. *Behavioral Brain Research*, *93*, 185–190. doi:10.1016/S0166-4328(98)00019-9

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence (Ed.). New York, NY: Academic Press.

Azvolinsky, A. (2018). A retracted paper on sex differences ignites debate. Retrieved May 3, 2020, from https://www.the-scientist.com/news-opinion/a-retracted-paper-on-sex-differences-ignites-debate-64873

Baddeley, A. (2018). *Exploring working memory. selected works of alan baddeley*. London: Routledge.

Bailey, J. M., Vasey, P. L., Diamond, L. M., Breedlove, S. M., Vilain, E., & Epprecht, M. (2016). Sexual orientation, controversy, and science. *Psychological Science in the Public Interest*, *17*(2), 45–101. doi:10.1177/1529100616637616

Baker, M., & Penny, D. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*, 452–454.

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533. doi:10.1038/483531a

Berenbaum, S. A., & Beltz, A. M. (2016). How early hormones shape gender development. *Current Opinion in Behavioral Sciences*, *7*, 53–60. Development and behavior. doi:10.1016/j.cobeha.2015.11.011

Bettencourt, B. A., & Miller, N. (1996). Gender differences in aggression as a function of provocation: A meta-analysis. *Psychol Bull*, *119*(3), 422–447.

Bonsang, E., Skirbekk, V., & Staudinger, U. M. (2017). As you sow, so shall you reap: Gender-role attitudes and late-life cognition. *Psychological Science*, *28*(9), 1201–1213. doi:10.1177/0956797617708634

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., … Zuber, o. b. o. t. S. C. C. T., Sabrina. (2013). Data Resource Profile: The Survey of Health, Ageing and Retirement in Europe (SHARE). *International Journal of Epidemiology*, *42*(4), 992–1001. doi:10.1093/ije/dyt088

Bostrom, N. (2014). *Superintelligence: Paths, dangers, strategies*. Oxford, United Kingdom: Oxford University Press.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433–1436. doi:10.1126/science.aaf0918

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Conway, M. A. (2005). Memory and the self. *Journal of Memory and Language*, *53*(4), 594–628. doi:10.1016/j.jml.2005.08.005

Conway, M. A., & Pleydell-pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review*, 261–288. doi:10.1037/0033-295X.107.2.261

Costa, P. T., Terracciano, A., & McCrae, R. R. (2001). Gender differences in personality traits across cultures: Robust and surprising findings. *Journal of Personality and Social Psychology*, *81*(2), 322–331. doi:10.1037/0022-3514.81.2.322

Deary, I. (2003). Population sex differences in IQ at age 11: The Scottish mental survey 1932. *Intelligence, 31*(6), 533–542. doi:10.1016/s0160-2896(03)00053-9

Delis, D. C., Kramer, J. H., Kaplan, E., & Ober, B. A. (1987). *Cvlt, california verbal learning test: Adult version: Manual*. Psychological Corporation.

Durlak, J. (2009). How to select, calculate, and interpret effect sizes. *Journal of pediatric psychology*, *34*(9), 917–928. doi:10.1093/jpepsy/jsp004

Eagly, A. H. (1995). The science and politics of comparing women and men. *American Psychologist*, *50*(3), 145–158. doi:10.1037/0003-066X.50.3.145

Ellis, H. (1894). *Man and woman*. London: Walter Scott.

Else-Quest, N. M., Hyde, J. S., & Linn, M. C. (2010). Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin*, *136*(1), 103–127. doi:10.1037/a0018053

Feingold, A. (1992). Cumulation of variance ratios. *Review of Educational Research*, *62*(4), 433–434.

Fillingim, R. B., King, C. D., Ribeiro-Dasilva, M. C., Rahim-Williams, B., & Riley, J. L. (2009). Sex, gender, and pain: A review of recent clinical and experimental findings. *The Journal of Pain*, *10*(5), 447–485. doi:10.1016/j.jpain.2008.12.001

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101(2)*, 171–191.

for Research in Education, S. C. (1933). *The intelligence of scottish children: A national survey of an the intelligence of Scottish children: A national survey of an age-group*. Univeristy of London Press.

for Research in Education, S. C. (1949). *The trend of Scottish intelligence.* Univeristy of London Press.

Fouchier, R. A. M., Garćıa-Sastre, A., Kawaoka, Y., Barclay, W. S., Bouvier, N. M., Brown, I. H., … Webster, R. G. (2013). Transmission studies resume for avian flu. *Science*, *339*(6119), 520–521. doi:10.1126/science.1235140

Francis, G. (2012). Too good to be true: Publication bias in two prominent studies from experimental psychology. *Psychonomic Bulletin & Review*, *19*(2), 151–156. doi:10.3758/s13423-012-0227-9

Frank, S. A. (2009). The common patterns of nature. *Journal of evolutionary biology*, *22*(8), 1563–1585. doi:10.1111/j.1420-9101.2009.01775.x

Fryar, C. D., Gu, Q., Ogden, C. L., & Flegal, K. M. (2016). Anthropometric reference data for children and adults: United States, 2011–2014. *Vital and Health Statistics*, *3*(39).

Gaulin, S. J. (1992). Evolution of sex differences in spatial ability. *Yearbook of Physical Anthropology*, *35*, 125–151.

Grabowska, A. (2016). Sex on the brain: Are gender-dependent structural and functional differences associated with behavior? *Journal of Neuroscience Research*, *95*, 200–212. doi:10.1002/jnr.23953

Grijalva, E., Newman, D. A., Tay, L., Donnellan, M. B., Harms, P. D., Robins, R. W., & Yan, T. (2015). Gender differences in narcissism: A meta-analytic review. *Psychological Bulletin*, *141*(2), 261–310. doi:10.1037/a0038231

Grissom, R., & Kim, J. (2005). *Effect sizes for research: A broad practical approach*. doi:10.4324/9781410612915

Guiso, L., Monte, F., Sapienza, P., & Zingales, L. (2008). Culture, gender, and math. *Science*, *320*, 1164–1165. doi:10.1126/science.1154094

Haier, R. J., Jung, R. E., Yeo, R. A., Head, K., & Alkire, M. T. (2005). The neuroanatomy of general intelligence: Sex matters. *NeuroImage*, *25*(1), 320–327. doi:10.1016/j.neuroimage.2004.11.019

Harris, S. (2017). #73 - Forbidden knowledge. Podcast. Retrieved March 3, 2020, from https://samharris.org/podcasts/forbidden-knowledge/

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science*, *269*, 41–45. doi:10.1126/science.7604277

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. doi:10.1016/C2009-0-03396-0

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research synthesis methods*, *1*(1), 39–65. doi:10.1002/jrsm.5

Herfst, S., Schrauwen, E. J. A., Linster, M., Chutinimitkul, S., de Wit, E., Munster, V. J., … Fouchier, R. A. M. (2012). Airborne transmission of influenza a/h5n1 virus between ferrets. *Science*, *336*(6088), 1534–1541. doi:10.1126/science.1213362

Herlitz, A., Airaksinen, E., & Nordström, E. (1999). Sex differences in episodic memory: The impact of verbal and visuospatial ability. *Neuropsychology*, *13(4)*, 590–597. doi:10.1037/0894-4105.13.4.590

Herlitz, A., Nilsson, L.-G., & Bäckman, L. (1997). Gender differences in episodic memory. *Memory & Cognition*, *25(6)*, 801–811. doi:10.3758/BF03211324

Herlitz, A., & Rehnman, J. (2008). Sex differences in episodic memory. *Current Directions in Psychological Science*, *17(1)*, 52–56. doi:10.1111/j.1467-8721.2008.00547.x

Herrnstein, R. J., & Murray, C. (1996). *The bell curve*. Free Press.

Hill, T. P. (2017). An evolutionary theory for the variability hypothesis. arxiv:1703.04184.

Hill, T. P. (2018). Academic activists send a published paper down the memory hole. Retrieved September 30, 2010, from https://quillette.com/2018/09/07/academic-activists-send-a-published-paper-down-the-memory-hole/

Hoare, C. (1980, October 27). The 1980 ACM Turing Award lecture. Web page. Retrieved February 27, 2020, from https://www.cs.fsu.edu/~engelen/courses/COP4610/hoare.pdf

Holcomb, H. R. (1996). Just so stories and inference to the best explanation in evolutionary psychology. *Minds and Machines*, *6*(4), 525–540. doi:10.1007/BF00389657

Hripsime, K. A., & Raudenbush, S. (1996). A multivariate mixed linear model for meta-analysis. *Psychological Methods*, *1*(3), 227–235. doi:10.1037/1082-989X.1.3.227

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability. *Psychological Bulletin*, *104(1)*, 53–69.

Iliescu, D., Ilie, A., Ispas, D., Dobrean, A., & Clinciu, A. I. (2016). Sex differences in intelligence: A multi-measure approach using nationally representative samples from romania. *Intelligence*, *58*, 54–61. doi:https://doi.org/10.1016/j.intell.2016.06.007

Ingalhalikar, M., Smith, A., Parker, D., Satterthwaite, T. D., Elliott, M. A., Ruparel, K., … Verma, R. (2014). Sex differences in the structural connectome of the human brain. *Proceedings of the national academy of sciences of the United States of America*, *111*(2), 823–828. doi:10.1073/pnas.1316909110

Ioannidis, J. P. A. (2005a). Contradicted and initially stronger effects in highly cited clinical research. *JAMA*, *294*(2), 218–228. doi:10.1001/jama.294.2.218

Ioannidis, J. P. A. (2005b). Why most published research findings are false. *PLOS medicine*, *2*, e124. doi:10.1371/journal.pmed.0020124

Irwing, P., & Lynn, R. (2005). Sex differences in means and variability on the progressive matrices in university students: A meta-analysis. *British Journal of Educational Psychology*, *96*, 505–524. doi:10.1348/000712605X53542

Jacobs, L. F., Gaulin, S. J., Sherry, D. F., & Hoffman, G. E. (1990). Evolution of spatial cognition: Sex-specific patterns of spatial behavior predict hippocampal size. *Proceedings of the National Academy of Sciences*, *87*(16), 6349–6352. doi:10.1073/pnas.87.16.6349

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. PMID: 22508865. doi:10.1177/0956797611430953

Johnson, W., Carothers, A., & Deary, I. J. (2008). Sex differences in variability in general intelligence: A new look at the old question. *Perspectives on Psychological Science*, *3*(6), 518–531. PMID: 26158978. doi:10.1111/j.1745-6924.2008.00096.x

Jonas, E., & Kording, K. P. (2017). Could a neuroscientist understand a microprocessor? *PLOS Computational Biology*, *13*(1), 1–24. doi:10.1371/journal.pcbi.1005268

Jones, C. M., Braithwaite, V. A., & Healy, S. D. (2003). The evolution of sex differences in spatial ability. *Behavioral Neuroscience*, *117*(3), 403–411. doi:10.1037/0735-7044.117.3.403

Jordan, K., Wüstenberg, T., Heinze, H. J., Peters, M., & Jäncke, L. (2002). Women and men exhibit different cortical activation patterns during mental rotation tasks. *Neuropsychologia*, *40*(13), 2397–408. doi:10.1016/S0028-3932(02)00076-3

Kleiman, D. G. (1977). Monogamy in mammals. *The Quarterly Review of Biology*, *52*(1), 39–69. PMID: 857268. doi:10.1086/409721

Konstantopoulos, S. (2011). Fixed effects and variance components estimation in three-level meta-analysis. *Research Synthesis Methods*, *2*(1), 61–76. doi:10.1002/jrsm.35

Kowal, P., Chatterji, S., Naidoo, N., Biritwum, R., Fan, W., Lopez Ridaura, R., … the SAGE Collaborators. (2012). Data Resource Profile: The World Health Organization Study on global AGEing and adult health (SAGE). *International Journal of Epidemiology*, *41*(6), 1639–1649. doi:10.1093/ije/dys210

Kret, M., & De Gelder, B. (2012). A review on sex differences in processing emotional signals. *Neuropsychologia*, *50*(7), 1211–1221. doi:10.1016/j.neuropsychologia.2011.12.022

Lakin, J. M. (2013). Sex differences in reasoning abilities: Surprising evidence that male–female ratios in the tails of the quantitative reasoning distribution have increased. *Intelligence*, *41*(4), 263–274. doi:https://doi.org/10.1016/j.intell.2013.04.004

Larsson, M. (2011). *A critique of some assumptions underlying scientific theories of consciousness, exemplified through a discussion of the integrated information theory of consciousness* (Master's thesis, University of Oslo).

Lehre, A.-C., Lehre, K. P., Laake, P., & Danbolt, N. C. (2009). Greater intrasex phenotype variability in males than in females is a fundamental aspect of the gender differences in humans. *Developmental Psychobiology*, *51*(2), 198–206. doi:10.1002/dev.20358

Lewin, C., Wolgers, G., & Herlitz, A. (2001). Sex differences favoring women in verbal but not in visuospatial episodic memory. *Neuropsychology, 15(2)*, 165–173. doi:10.1037//0894-4105.15.2.165

Lippa, R. A., Collaer, M. L., & Peters, M. (2010). Sex differences in mental rotation and line angle judgments are positively associated with gender equality and economic development across 53 nations. *Archives of Sexual Behavior*, *39*(4), 990–997. doi:10.1007/s10508-008-9460-8

Lohman, D. F., & Lakin, J. M. (2009). Consistencies in sex differences on the Cognitive Abilities Test across countries, grades, test forms, and cohorts. *British Journal of Educational Psychology*, *79*, 389–407. doi:10.1348/000709908X354609

Maccoby, E. E., & Jacklin, C. N. (1975). *The psychology of sex differences*. Stanford: Stanford University Press.

Machin, S., & Pekkarinen, T. (2008). Global sex differences in test score variability. *Science*, *322*(5906), 1331–1332. doi:10.1126/science.1162573

Maylor, E. A., Reimers, S., Choi, J., Collaer, M. L., Peters, M., & Silverman, I. (2007). Gender and sexual orientation differences in cognition across adulthood: Age is kinder to women than to men regardless of sexual orientation. *Archives of Sexual Behavior, 36(2)*, 235–249. doi:10.1007/s10508-006-9155-y

Mazei, J., Hüffmeier, J., Freund, P. A., Stuhlmacher, A. F., Bilke, L., & Hertel, G. (2015). A meta-analysis on gender differences in negotiation outcomes and their moderators. *Psychological Bulletin*, *141*(1), 85–104. doi:10.1037/a0038184

McCarthy, M. M. (2016). Multifaceted origins of sex differences in the brain. *Philosophical Transactions B*, *371*, 1–11. doi:10.1098/rstb.2015.0106

Mcgrath, B., & Meyer, G. (2006). When effect sizes disagree: The case of r and d. *Psychological methods*, *11*(4), 386–401. doi:10.1037/1082-989X.11.4.386

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, *67*(4), 371–378. doi:10.1037/h0040525

Moore, M. C. (1991). Application of organization-activation theory to alternative male reproductive strategies: A review. *Hormones and Behavior*, *25*(2), 154–179. doi:https://doi.org/10.1016/0018-506X(91)90048-M

Nakagawa, S., Poulin, R., Mengersen, K., Reinhold, K., Engqvist, L., Lagisz, M., & Senior, A. M. (2015). Meta-analysis of variation: Ecological and evolutionary applications and beyond. *Methods in*

Ecology and Evolution, *6*(2), 143–152. doi:10.1111/2041-210X.12309

Neuroskeptic. (2012). The nine circles of scientific hell. *Perspectives on psychological science : a journal of the Association for Psychological Science*, *7*(6), 643–644.

Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance, and extreme scores. *Sex Roles*, *39*(1), 21–43. doi:10.1023/A:1018873615316

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). doi:10.1126/science.aac4716

Petersen, J. L., & Hyde, J. S. (2010). A meta-analytic review of research on gender differences in sexuality, 1993–2007. *Psychological Bulletin*, *136*(1), 21–38. doi:10.1037/a0017504

Pietschnig, J., & Voracek, M. (2015). One century of global iq gains: A formal meta-analysis of the flynn effect. *Perspectives on Psychological Sciences*, *10(3)*, 282–306. doi:10.1177/1745691615577701

Proverbio, A. M. (2017). Sex differences in social cognition: The case of face processing. *Journal of neuroscience research*, *95*(1-2), 222–234. doi:10.1002/jnr.23817

Ragland, J., Coleman, A., Gur, R., Glahn, D., & Gur, R. (2000). Sex differences in brain-behavior relationships between verbal episodic memory and resting regional cerebral blood flow. *Neuropsychologia*, *38*(4), 451–461. doi:10.1016/S0028-3932(99)00086-X

Re, A. C. D. (2013). *compute.es: Compute effect sizes*. Retrieved from http://cran.r-project.org/web/packages/compute.es

Reder, L. M. (1996). *Implicit memory and metacognition*. Psychology Press.

Reinhold, K., & Engqvist, L. (2013). The variability is in the sex chromosomes. *Evolution*, *67*(12), 3662–3668. doi:10.1111/evo.12224

Rosenfield, S., & Mouzon, D. (2013). Gender and mental health. In C. S. Aneshensel, J. C. Phelan, & A. Bierman (Eds.), *Handbook of the sociology of mental health* (pp. 277–296). doi:10.1007/978-94-007-4276-5_14

Russell, C. A., Fonville, J. M., Brown, A. E. X., Burke, D. F., Smith, D. L., James, S. L., … Smith, D. J. (2012). The potential for respiratory droplet–transmissible a/h5n1 influenza virus to evolve in a mammalian host. *Science*, *336*(6088), 1541–1547. doi:10.1126/science.1222526

Savolainen, V., & Hodgson, J. A. (2016). Evolution of homosexuality. In V. Weekes-Shackelford, T. K. Shackelford, & V. A. Weekes-Shackelford (Eds.), *Encyclopedia of evolutionary psychological science* (pp. 1–8). doi:10.1007/978-3-319-16999-6{_}3403-1

Scammacca, N., Roberts, G., & Stuebing, K. K. (2014). Meta-analysis with complex research designs: Dealing with dependence from multiple measures and multiple group comparisons. *Review of Educational Research*, *84*, 328–364. doi:10 . 3102 / 0034654313500826

Schacht, R., & Bell, A. V. (2016). The evolution of monogamy in response to partner scarcity. *Scientific reports*, *6*, 32472–32472. doi:10.1038/srep32472

Sonnega, A., Faul, J. D., Ofstedal, M. B., Langa, K. M., Phillips, J. W., & Weir, D. R. (2014). Cohort Profile: the Health and Retirement Study (HRS). *International Journal of Epidemiology*, *43*(2), 576–585. doi:10.1093/ije/dyu067

Sperling, G. (1960). The information available in brief visual presentations. *Psychological monographs*, *74*.

Steptoe, A., Breeze, E., Banks, J., & Nazroo, J. (2012). Cohort Profile: The English Longitudinal Study of Ageing. *International Journal of Epidemiology*, *42*(6), 1640–1648. doi:10.1093/ije/dys168

Stoet, G., & Geary, D. (2019). A simplified approach to measuring national gender inequality. *PLOS ONE*, *14*, e0205349. doi:10.1371/journal.pone.0205349

Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within- and across-nation assessment of 10 years of PISA data. *PLOS ONE*, *8*(3), e57988. doi:10.1371/journal.pone.0057988

Stoet, G., & Geary, D. C. (2015). Sex differences in academic achievement are not related to political, economic, or social equality. *Intelligence*, *48*, 137–151. doi:10.1016/j.intell.2014.11.006

Strand, S., Deary, I. J., & Smith, P. (2006). Sex differences in cognitive abilities test scores: A UK national picture. *Brittish Journal of Educational Psychology*, *76*, 463–480. doi:10 . 1348 / 000709905X50906

Strandqvist, A. (2018). *Psychological aspects in differences/disorders of sex development* (Doctoral dissertation, Karolinska Institutet).

The PLOS ONE Staff. (2019). Correction: The magnitude of sex differences in verbal episodic memory increases with social progress:

Data from 54 countries across 40 years. *PLOS ONE*, *14*(5), 1–3. doi:10.1371/journal.pone.0217033

The World Bank. (n.d.). *Databank*. Retrieved from http://databank.worldbank.org/data/home.aspx

Tian, L., Wang, J., Yan, C., & He, Y. (2011). Hemisphere- and gender-related differences in small-world brain networks: A resting-state functional MRI study. *NeuroImage*, *54*(1), 191–202. doi:10.1016/j.neuroimage.2010.07.066

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory*. New York: Academic Press.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology*, *53*, 1–25. doi:10.1146/annurev.psych.53.100901.135114

United Nations Development Programme. (n.d.). *Human development reports*. Retrieved from http://hdr.undp.org/en/data

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. Retrieved from http://www.jstatsoft.org/v36/i03/

Voyer, D. (2011). Time limits and gender differences on paper-and-pencil tests of mental rotation: A meta-analysis. *Psychological Bulletin, 18(2)*, 267–277. doi:10.3758/s13423-010-0042-0

Voyer, D., Postma, A., Brake, B., & Imperato-McGinley, J. (2007). Gender differences in object location memory: A meta-analysis. *Psychonomic Bulletin and Review, 14(1)*, 23–38. doi:10.3758/s13423-010-0042-0

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin, 140(4)*, 1174–1204. doi:10.1037/a0036620

Voyer, D., Voyer, S. D., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. *Psychological Bulletin, 117(2)*, 250–270.

Weber, D., Dekhtyar, S., & Herlitz, A. (2017). The Flynn effect in Europe — Effects of sex and region. *Intelligence*, *60*(1), 39–45.

Weber, D., Skirbekk, V., Freund, I., & Herlitz, A. (2014). The changing face of cognitive gender differences in europe. *Proceedings of the National Academy of Sciences*, *111*(32), 11673–11678. doi:10.1073/pnas.1319538111

Wierenga, L. M., Doucet, G. E., Dima, D., Agartz, I., Aghajani, M., Akudjedu, T. N., … Tamnes, C. K. (2020). Greater male than female variability in regional brain structure across the lifespan. *bioRxiv*. doi:10.1101/2020.02.17.952010

Williams, R. L. (2013). Overview of the flynn effect. *Intelligence, 41(6)*, 753–794. doi:10.1016/j.intell.2013.04.010

World Economic Forum. (2020). *Global gender gap report 2020*. Retrieved from https://www.weforum.org/reports/gender-gap-2020-report-100-years-pay-equality

World Health Organization. (2020). *Who director-general's opening remarks at the media briefing on covid-19 - 11 march 2020*. Retrieved March 11, 2020, from https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020

Young, K. D., Bellgowan, P. S., Bodurka, J., & Drevets, W. C. (2013). Functional neuroimaging of sex differences in autobiographical memory recall. *Human Brain Mapping, 34*(12), 3320–3332. doi:10.1002/hbm.22144

Zakzanis, K. K. (2001). Statistics to tell the truth, the whole truth, and nothing but the truth: Formulae, illustrative numerical examples, and heuristic interpretation of effect size analyses for neuropsychological researchers. *Archives of Clinical Neuropsychology, 16*(7), 653–667. doi:10.1016/S0887-6177(00)00076-7

Zerjal, T., Xue, Y., Bertorelle, G., Wells, R. S., Bao, W., Zhu, S., … Tyler-Smith, C. (2003). The genetic legacy of the mongols. *American journal of human genetics, 72*(3), 717–721. doi:10.1086/367774