LEX

Luxemburger Experimental-praktikum Journal

Zeitschrift psychologischer Forschung
Revue de recherche en psychologie

Band 16, Jahrgang 2022

Université du Luxembourg
Bachelor scientifique en psychologie

ISSN 3093-1045

Luxemburger Experimentalpraktikum Journal
Band 16, Jahrgang 2022

Inhalt

Rapid automatized Naming (RAN) in younger and older adults

Amelie I. Buttkus, Alicia S. Funk, Julie F. Neuenfeldt, Leon A. Orlik, Simon P. Schommer, Estelle Spizzica

Supervisor: Dr. Caroline Hornung

Rapid automatized naming (RAN) is defined as the ability to name familiar visual stimuli such as colors, letters and digits as fast as possible. Although RAN has been used to predict adult reading ability, there have been few applications for mathematical abilities. In the present study, the main goal was to focus on differences in sixty-four younger and older adults in RAN and arithmetic performance in order to report possible changes throughout adulthood. Six RAN tasks were administrated consisting of digit, dice, canonical finger configuration and newly introduced roman number, dot and non-canonical finger configuration. The standardized arithmetic test "Tempo Test Rekenen" (De Vos, 1992) was administrated to measure participants' arithmetic performance, subdivided into five columns of basic arithmetic operations. The results underline that younger adults perform significantly better in RAN and arithmetic tasks than older adults. The newly introduced RAN tasks appeared to be a good complement to the existing ones. It is assumed that RAN is a predictor for arithmetic fluency in adults.

1. Introduction

described as the ability to name visual stimuli such as colors, letters, or digits as fast as possible (Hornung et al., 2017; Koponen et al., 2016). Since a better part of existing literature focuses on the relation between RAN and reading abilities, the tasks have turned out as a valid predictor of reading skills and deficits such as dyslexia in school-aged children (Koponen et al., 2016; Lervåg & Hulme, 2009). Denckla and Cutting (1999) traced the origins of RAN in the description of a neurological phenomenon called "alexia without agraphia" back in the nineteenth century. This depiction is based on a paper by Geschwind and Fusillo (1966) who observed cases of adults with brain lesions that resulted in a disruption of the "visual-verbal" connection (Denckla & Cutting, 1999). This means that the patients lose their ability to read while writing remains possible. During their work, Geschwind and Fusillo detected difficulties in the correct naming of colors as visual stimuli in their patients, administered as a recovery measure for adults with head injuries resulting in the loss of reading skills. The assumed relation between reading skills and the naming of visual stimuli led them to examine first graders with reading difficulties in their capacities of naming visual stimuli. These pupils were indeed able to name the colors, but they showed slow reaction times and a lack of automatism. This observed relation is the origin of the development of RAN-tasks and the effort to validate RAN as a predictor of reading skills.

As confirmed by later research, RAN and reading do share related cognitive processes (Hornung et al., 2017). Those refer to the visual-verbal connection and executive functions. These shared cognitive functions are also relevant for processing speed, which is important for reading and RAN as well, and which is measured with RAN tasks (Denckla & Cutting, 1999).

RAN has been a very useful measure of cognitive abilities not only for children, but it has been proven for younger adults as well (Gordon et al., 2021). Although RAN has been used to predict adult reading ability, there have been few applications for mathematical abilities (Wiig

et al., 2002). Additionally, the effect of aging on RAN is less well described.

Prior studies distinguished symbolic RAN tasks using digits or letters and non-symbolic RAN tasks using stimuli like colors (Gordon et al., 2021).

Furthermore, the different types of stimuli can also be distinguished between alphanumeric stimuli (written symbols) and non-alphanumeric stimuli (not written symbols, such as colors, objects, dots, dice (Koponen et al., 2017)). There are different purposes for the different types of tasks (Koponen et al., 2017). Alphanumeric RAN shows a stronger relation with reading skills of children than non-alphanumeric RAN. That's different regarding arithmetic. Although both stimuli types described similarly relation to arithmetic (e.g., Koponen et al., 2013 in Hornung et al., 2017), non-alphanumeric RAN is assumed to be a valid predictor for arithmetic fluency (Hornung et al., 2017).

In general, performance varies for different RAN tasks. Usually, performances in alphanumeric RAN are faster than in non-alphanumeric. But through the explicit learning of numbers, experience is gained with these symbolic stimuli and therefore the performance in symbolic RAN tasks of adults is better than in non-symbolic RAN tasks (Koponen et al., 2017; Gordon et al., 2021). One explanation is that serial naming of digits and letters is more automatized than naming of colors or objects, which demands further additional cognitive processes such as attributing meaning (Roelofs, 2006 in Hornung et al., 2017).

RAN itself can be seen as an index of speed of lexical access and the quick retrieval of phonological representations. It measures the processing speed, which can help to evaluate cognitive functions underlying recognition, memory, reading and language production (Jacobson et al., 2004). More importantly, RAN and arithmetic fluency tasks do share related cognitive capacities such as executive functions, processing speed, and the retrieval of phonological representations from long-term memory, such as number words. Here as well, explicit learning of digits is important as it automates fast retrieval and thus enables fast solving of single - digit arithmetic tasks as well as a better RAN performance (Koponen et al., 2017). This could also explain why there is no improvement in symbolic RAN during adulthood (36- 65 years) and no differences in performances of younger and older adults could be found in symbolic (letter and digit) RAN tasks. Although older adults showed significantly slower performances than younger adults in non-symbolic (color and object) RAN tasks (Gordon et al., 2021). However, Jacobson et al. (2004) found a significant relationship between age and naming time for both types of RAN tasks in a sample of 15 to 85 years old participants. Time would be expected to slow by about 1 sec. every 25 years for different types of RAN tasks.

It is important to provide a general overview of arithmetic problem solving, and the related cognitions to better understand the relation between RAN and arithmetic.

Differences between young and older adults in arithmetic problem solving, regarding speed and accuracy were found. In both aspects, older adults performed worse. These differences increased with complexity of the problems (Duverne and Lemaire, 2005; Uittenhove and Lemaire, 2014, as cited in Hinault and Lemaire, 2016).

To achieve an improved understanding of the decline of arithmetic problem-solving abilities in older adults, it is essential to amplify which cognitions are influencing those changes in age. Furthermore, many of the studies on age-related changes in arithmetic abilities concentrate on strategies differences in younger and older adults, as they are connected.

Hinault and Lemaire (2016) elaborated that the executive control processes, more precisely inhibition, cognitive flexibility and working memory, are influencing for example the strategy use. Older adults employ less strategies than younger adults, which can be explained by a decrease in inhibition of irrelevant answers and cognitive flexibility resources switching between different cognitive concepts.

The function of the working memory is to store information for a short period of time and processing the information to execute cognitive tasks (Imbo & Vandierendonck, 2007).

Moreover, there is a shift in older adults to strategy-use involving less executive cognitive processes like arithmetic fact retrieval (Hinault &

Lemaire, 2016). This strategy draws the solution directly from long-term memory (Thevenot et al., 2007). The authors interpreted the more frequent use of retrieval by older adults with a longer practical experience with arithmetic problems and the resulting expansion and enrichment of their fact network. This could also explain why older adults perform as well as younger adults in particular arithmetic tasks (Hinault & Lemaire, 2016). Adults apply the retrieval strategy for example in simple addition problems (Thevenot et al., 2007). However, it was also discovered that older adults take longer to find the answer in their fact network (Hinault & Lemaire, 2016). Imbo and Vandierendonck (2007) explained this result by the fact that executive working memory resources are influencing the retrieval process. They further elaborated that addition and subtraction employ 88% and 72% of retrieval. In these mathematical operations, the process of counting (non-retrieval) is often applied (Imbo & Vandierendonck, 2007). Multiplication uses 98% retrieval and division 69%. Both are reliant on retaining the answer to the corresponding problem. In contrast to multiplication, division problems require a more considerable use of the executive working memory (Imbo & Vandierendonck, 2007).
Overall, 83% of the procedures used in addition with sums lower than 10 (small numbers) were accounted for retrieval and 46% in operations with sums greater than 10 (medium numbers). In that study, most of the adults used retrieval strategies for small numbers and non-retrieval strategies for medium numbers (LeFevre et al., 1996 as cited in Thevenot et al., 2007). The non-retrieval strategies like transformation or counting must also retrieve arithmetic facts from long-term memory, but also require other cognitive processes like reciting numbers, calculating, and storing intermediate results. For the latter, the phonological working memory is essential. However, it is also supposed that phonological processes are relevant in multiplication operations because multiplication makes use of phonological codes (Imbo & Vaniderendonck, 2007; Moeller et al., 2011). De Smedt and Boets (2010) could prove that phonological processing, more precisely phonological awareness (“the ability to recognize and distinguish between the sounds used in spoken language” (APA, 2020)) and arithmetic fact retrieval correlate in adults and that this relation was higher for multiplication compared to subtraction. These different results suggest that the function of the specific areas of phonological processing for arithmetic problem solving should be further investigated.

1.1. Present study

As mentioned before, RAN and arithmetic show significant correlations, which is why RAN can be considered as a predictor for arithmetic fluency in children (Koponen et al., 2016). Since these characteristics have only been examined for school-aged children, we want to explore if this relation also exists for adults. Moreover, in children, the correlations between arithmetic fluency and number-specific RAN tasks turned out to be stronger than with other forms of RAN (Hornung et al., 2017). Since we aim to examine the relation between arithmetic tasks and naming speed of arithmetic depictions, we only used numeric RAN tasks and did not apply any other RAN tasks as color or objects, which are normally used as well. Furthermore, we introduced three new RAN-tasks. As a new symbolic task, we used roman numbers and as new non-symbolic RAN tasks we used arrays of 1 to 5 dots and non-canonical fingers. We want to test whether these tasks are suitable as a complement for the existing RAN tasks.
The main goal of our research is to find out whether there are any differences between younger and older adults regarding their capacities in RAN and arithmetic in order to report possible changes throughout adulthood. First of all, we expect younger adults to be faster in RAN-tasks than older adults. Given that several executive functions are relevant for mental tasks and naming times degrade with age (Jacobson et al., 2004), we expect younger participants to achieve faster results.
Our second hypothesis is that we assume younger adults to solve more arithmetic tasks, given that older adults tend to perform worse in arithmetic problem solving (Hinault & Lemaire, 2016). This difference between the age groups

is only expected in the addition and subtraction tasks. These tasks rely on executive functions as well. As division and multiplication tasks are more similar to an automatized retrieval of learned information, the cognitive processes which pass while doing simple tasks in these two calculation types do not correspond to proper calculation (Imbo & Vandierendonck, 2007). In consideration of the fact that these processes are not expected to decline in their functionality with aging (Radvansky et al., 1996), we expect the younger age group to not show significantly better results than people of the older age group.

As a third hypothesis in regard to the relation between RAN and arithmetic, we suppose that with a faster naming time of a participant, the more arithmetic tasks will be solved correctly. Thus, we expect the predictor effect of RAN for arithmetic fluency to exist in adults as well, regardless of their age group.

Lastly, we hypothesized that symbolic RAN-tasks will show a higher correlation with arithmetic in both age groups than non-symbolic RAN. To predict arithmetic fluency in children, non-symbolic RAN can be applied (Hornung et al., 2017), but we expect changes in this relation in adults because of the amount of practice of symbolic stimuli. (Gordon et al., 2021; Koponen et al., 2016).

2. Method

2.1. Participants

The sample of the study comprised 64 participants, divided into two independent subgroups. The first subgroup consisted of 33 young adults aged between 16 and 30 years, whereby 16 participants were female and 17 were male. The average age of the younger subgroup was M = 21.03 (SD = 3.27) years. The second subgroup consisted of 31 elderly adults aged between 60 and 85 years, whereby 13 participants were male and 18 were female. The average of the elderly subgroup was M = 67.65 (SD = 5.52) years. The average age at large was M = 43.61 (SD = 23.90) years.

The study included people from various occupational groups and with different qualification levels. The majority of the study participants had a general qualification for university entrance (N = 22) as the highest qualification. 21 participants had a completed vocational training. Slightly fewer, 13 participants had a completion of a university degree.

According to this, most of the participants were students (N = 19), which was also visible in the occupational group section. The most frequently represented occupational group after the student selection option, was service occupations and salespersons, with a quantity of 12 people.

For a distinct majority of 84.4% of the participants, the native language was German. Approximately 71% of the participants indicated, that they do not speak any second language, while at least 18.8% stated French as their second language. In addition, 81.3% of the 64 participants were right-handed while the rest were left-handed.

An exclusion criterion to participate in the study was being diagnosed with Dyscalculia. There were no participants who were being diagnosed with Dyscalculia in the past, taking part in the study.

All participants had normal or corrected-to-normal visual acuity and have been informed about the purpose of the study.

2.2. Measures

Rapid automatized naming tasks (RAN). First of all, six RAN tasks were administrated using five recurring “quantities” ranging from 1 to 5 represented in digit form, in roman number form, in dice form, in canonical finger configuration, in non-canonical finger configuration and in dot form. Both finger RAN tasks were inspired from Hornung et al. (2017). To our knowledge, the roman number, the dots-RAN and non-canonical fingers RAN-tasks have never been used in prior research.

The stimuli of the test were randomly arrayed in five rows of eight resulting in a total number of 40 stimuli per RAN task. Participants had to name the stimuli as fast as they can without doing any mistakes. A stopwatch was used to measure the time. The number of mistakes by naming a false number was taken into account.

In total we had a very low mean error rate of 0.49%.
Each of these RAN tasks was conducted twice, with a different arrangement of the stimuli over both trials. By doing so, we also aimed to avoid possible memory effects of the participants. This excludes the possibility that the first trial was merely memorized and recited in the second one. We controlled the order of presentation of the RANs by rotating the successive RAN tasks, to prevent a possible bias in administration order. One administration order was for instance the naming of dice, digits, canonical finger patterns, roman numbers, non-canonical finger patterns and dot patterns whereas the next administration order started with the naming of digits and ended with the naming of dice patterns and so on. Before recording reaction time of the participants, they were asked to practice and name a row of five stimuli corresponding to the following RAN task. Between each task, participants had 15 seconds to rest. After practice, reaction times were recorded on each RAN measure. For each RAN task, the mean reaction times were calculated over both trials and used as mean RAN task performance in the statistical analyses.
A test-retest reliability analysis was conducted to determine the reliability of the RAN tasks used in the study. Pearson's correlation coefficient showed an excellent reliability between the two test runs, $r = .98$ for all tasks confounded. A more detailed retest-reliability analysis revealed that the correlation between both time points within each task was significantly high [$r = .89$; $r = .96$], $p = .00$.
A reliability analysis was conducted as well to determine the internal consistency of all six measures of the RAN tasks. The internal consistency was also excellent, $\alpha = .95$.

Arithmetic Skills (TTR). The standardized arithmetic test "Tempo Test Rekenen" (De Vos, 1992) was administrated to measure participants' arithmetic performance. The test was subdivided into five columns of 40 items each, resulting in a total number of 200 items. Four columns consisted of basic arithmetic operations such as addition, subtraction, multiplication and division. In the fifth column participants were asked to solve a mixture of those four basic arithmetic operations. The items continuously increase in difficulty from single digit to multidigit operations. The solutions of the items did not exceed 99. Participants were requested to solve the tasks as correct and as fast as possible within one minute. Between each column participants had 15 seconds to rest. A reliability analysis was conducted to determine the internal consistency of the items of the Tempo Test Rekenen. The internal consistency level was excellent, $\alpha = .927$. Corresponding to the excellent levels of reliability and internal consistency, it can be noted that the RAN tasks and the TTR test fulfilled the quality criterion of reliability.

Questionnaire. At the end of the test session, participants were asked to fill in a questionnaire with items about general demographic information (e.g., language and socioeconomic background) and items investigating frequencies of cognitive and free time activities such as reading, crossword puzzles and physical activities.

2.3. Procedure

Participants were tested individually in a quiet room inside their home by the same test administrator. All tasks were paper pencil. The participants first completed the six RAN tasks and then the arithmetic tasks. The test session lasted about 20 minutes.
Informed consent was obtained for all participants and their parents' participants were minors (n = 2). The study was conducted in compliance with national and European ethical norms related to research with human participants.

2.4. Statistical analysis

All data analyses were performed with SPSS. To investigate our hypotheses, we looked specifically at descriptive analyses such as mean values and standard deviations, t-tests, one-way MANOVA and correlations of the data collected.

All data was checked for outliers by conducting an explorative data analysis, but no such were found.
More in detail, two independent t-tests were performed to find differences between the two age groups regarding the tested variables i.e., performance in RAN and arithmetic. Errors have not been evaluated and we will not continue to work with them because the error rate was very low and they have been integrated into the response time, as mentioned above.
As for the correlation analyses, both Spearman and Pearson correlations were performed depending on the scale level of the respective variables treated. By performing correlation analyses, we aimed to test for relationships between the individual variables which have been collected either by the questionnaire or through the tasks performed.
The questionnaire collected among others, data on the leisure activities of our subjects. Out of interest, we looked at the correlations between these activities and the RAN and TTR tasks. Except for reading frequency and riddle solving, we had to further disregard data previously collected in the questionnaire, as we did not have enough representatives of the different response options to find meaningful results.

3. Results

As presented in Table 1, the younger age group was faster in performing the RAN tasks and showed more correct responses in the arithmetic tasks.
An independent t-test was conducted for the RAN performance in general. The Levene's test revealed that the variances of the two age groups were not equal, p = .02. The results show that the performance (i.e. faster reaction times) of the younger age group (M = 17.12, SD = 2.13) is significantly higher, $t(52) = -8.02$, p = .00, than the performance of the older age group (M = 22.51, SD = 3.13). We can state that the younger adults were faster in solving the RAN tasks.
An independent t-test was conducted for the arithmetic tasks in general and the Levene's test revealed that the variances of both age groups are equal. The results show that the performance of the younger age group (M = 29.06, SD = 5.60) is significantly higher than the performance of the older age group (M = 26.33, SD = 5.13), $t(62) = 2.03$, p = .047. We can suggest that the younger age group solved more correct arithmetic tasks than the older age group.
Table 2 showed the performance differences of both age groups in each RAN task. The conducted one-way MANOVA revealed, that both age groups were fastest in digit RAN and the slowest in the non-canonical finger task. However, the younger age group was faster in all RAN tasks, compared to the older age group.
A one-way MANOVA was conducted for the arithmetic tasks to analyze the difference between the performances of both age groups concerning addition, subtraction, multiplication, division and mixed arithmetic tasks. Levene's test revealed an equality of variances in all arithmetic tasks. We found that the younger age group shows a significantly higher performance in solving the addition (M = 33.27, SD = 3.65), $F(1, 62) = 7.96$, p = .006 and the subtraction tasks (M = 30.39, SD = 4.98), $F(1, 62)$ = 7.01, p = .010, than the older age group in the addition (M = 30.65, SD = 3.80) and the subtraction tasks (M = 27.03, SD = 5.19).
In regard to the multiplication and division tasks we found that in the multiplication tasks the younger age group (M = 26.30, SD = 6.98), $F(1, 62) = .03$, p = .853, does not differ significantly from the older age group (M =26.00, SD = 6.01) and the younger age group (M = 26.88, SD = 8.07), $F(1, 62) = 3.77$, p = .057, does also not differ significantly from the older age group (M = 22.77, SD = 8.83) in the division task. Additionally, we found that the younger age group (M = 28.45, SD = 6.10), $F(1, 62) = 5.05$, p = .028 performed significantly better in the mixed arithmetic task than the older age group (M = 25.19, SD = 5.46) (cf. Table 3).

Table 1
Descriptive statistics of RAN and arithmetic performances for both age groups.

	16-30 years	60-85 years

	N	*M*	*SD*	*N*	*M*	*SD*
RAN tasks	33	17.12	2.13	31	22.51	3.13
TTR tasks	33	29.06	0.98	31	26.33	0.92

Table 2
One-way MANOVA for RAN performances for both age groups.

	16-30 years (*N* = 33)		60-85 years (*N* =31)				
	M	*SD*	*M*	*SD*	*F*(1,62)	*p*	η^2
Non-canonical	20.52	2.64	28.20	3.45	100.77	<.001	.619
Digit	11.97	1.84	15.17	3.29	23.35	<.001	.274
Roman	16.72	2.63	20.64	4.90	16.15	<.001	.207
Dots	19.30	2.90	25.29	4.22	44.21	<.001	.416
Canonical	19.86	3.06	27.33	3.65	78.91	<.001	.560
Dice	14.31	2.17	18.44	2.67	46.46	<.001	.428

We ran a Pearson correlation to identify the correlation between participants' general RAN and arithmetic performance. The Pearson correlation coefficient was negative and highly significant (r = -.59, p = .00). In this case we can state that when a participant shows a better performance in the arithmetic task i.e., more right answers, he needed less time in the RAN tasks. Concerning the correlations within the RAN tasks, the Pearson correlation showed that the dots task correlated the highest with the composite score of the arithmetic tasks (r = -.63, p = .00). All in all, we only had moderate to high correlations between the RAN tasks and the arithmetic tasks, as shown in table 4.

We also looked at the correlations in both age groups in order to examine whether there were differences in the relationship between RAN and arithmetic regarding age. The Pearson correlation coefficient revealed for the younger age group a negative and very high correlation between the RAN tasks and the arithmetic tasks (r = -.70, p = .00). The Pearson correlation coefficient revealed for the older age group also a negative and high correlation (r = -.53, p = .003).

Table 5 shows, that in the younger age group, the arithmetic tasks correlate significantly the highest with the roman number RAN and the lowest with the non-canonical RAN.

In the older age group, the arithmetic tasks significantly correlated the highest with the dots-RAN and the lowest with the canonical RAN, as shown in table 6. Only the non-canonical RAN correlated even lower with the arithmetic tasks but not significantly.

Table 4
Pearson correlations between RAN and arithmetic (TTR) performances for both age groups.

		N	*M*	*SD*	1	2	3	4	5	6
1	Non-canonical	64	24.24	4.92	-					
	Digit	64	13.52	3.08	.447***	-				
	Roman	64	18.62	4.34	.489***	.761***	-			
	Dots	64	22.20	4.67	.565***	.594***	.707***	-		
	Canonical	64	23.48	5.02	.780***	.529***	.616***	.700***	-	
	Dice	64	16.31	3.18	.603***	.807***	.779***	.678***	.651***	-
	TTR	64	27.74	5.51	-.346**	-.543***	-.534***	-.634***	-.460***	-.550***

*Note: p < .05 *, p < .01 ** and p < .001 ***.*

Table 5
Pearson correlations between RAN and arithmetic (TTR) performances for the younger age group.

16-30 years		*N*	*M*	*SD*	1	2	3	4	5	6
1	Non-canonical	33	20.52	2.64	-					
2	Digit	33	11.97	1.84	.322	-				
3	Roman	33	16.72	2.63	.599***	.665***	-			
4	Dots	33	19.30	2.90	.492**	.653***	.693***	-		
5	Canonical	33	19.86	3.06	.742***	.454**	.663***	.651***	-	
6	Dice	33	14.31	2.17	.600***	.761***	.863***	.778***	.661***	-
7	TTR	33	29.06	5.60	-.421*	-.668***	-.691***	-.680***	-.455**	-.662***

*Note: p < .05 *, p < .01 ** and p < .001 ***.*

Table 6
Pearson correlations between RAN and arithmetic (TTR) performances for the older age group.

60-85 years		*N*	*M*	*SD*	1	2	3	4	5	6
1	Non-canonical	31	28.20	3.45	-					
2	Digit	31	15.17	3.29	.440*	-				
3	Roman	31	20.64	4.90	.442*	.789***	-			
4	Dots	31	25.29	4.22	.532**	.499**	.720***	-		
5	Canonical	31	27.33	3.65	.775***	.521**	.601***	.687***	-	
6	Dice	31	18.44	2.67	.570**	.849***	.765***	.599***	.612***	-
7	TTR	31	26.33	5.13	-.214	-.474**	-.475**	-.599***	-.415*	-.417*

*Note: p < .05 *, p < .01 ** and p < .001 ***.*

We went into further detail and investigated how the different RAN tasks correlate with the individual arithmetic operations in both age groups. Table 7 shows, that concerning the younger age group the roman number RAN and the addition task correlate significantly the highest. The canonical RAN and the subtraction task correlate significantly the lowest.

Table 7
Pearson correlations between RAN and individual arithmetic operations (TTR) performances for the younger age group.

16-30 years		1	2	3	4	5	6	7	8	9	10
1	Non-canonical	-									
2	Digit	.322	-								
3	Roman	.599***	.665***	-							
4	Dots	.492**	.653***	.693***	-						
5	Canonical	.742***	.454**	.663***	.651***	-					
6	Dice	.600***	.761***	.863***	.778***	.661***	-				
7	Addition	-.395*	-.624***	-.791***	-.638***	-.438*	-.662***	-			
8	Subtraction	-.395*	-.537**	-.601***	-.623***	-.393*	-.558**	.819***	-		
9	Multiplication	-.475**	-.677***	-.669***	-.696***	-.492**	-.664***	.853***	.823***	-	
10	Division	-.407*	-.661***	-.641***	-.679***	-.418*	-.634***	.828***	.863***	.916***	-
11	Mixed	-.293	-.608***	-.596***	-.536**	-.392*	-.589***	.770***	.855***	.853***	.868***

*Note: p < .05 *, p < .01 ** and p < .001 ***.*

Table 8
Pearson correlations between RAN and individual arithmetic operations (TTR) performances for the older age group.

60-85 years		1	2	3	4	5	6	7	8	9	10
1	Non-canonical	-									
2	Digit	.440*	-								
3	Roman	.422*	.789***	-							
4	Dots	.532**	.499**	.720***	-						
5	Canonical	.775***	.521**	.601***	.687***	-					
6	Dice	.570**	.849***	.765***	.599***	.612***	-				
7	Addition	-.473**	-.401*	-.343	-.419*	-.444*	-.479**	-			
8	Subtraction	-.074	-.342	-.358	-.454*	-.237	-.203	.599***	-		
9	Multiplication	-.191	-.499**	-.472**	-.564**	-.447*	-.378*	.570**	.717***	-	
10	Division	-.185	-.400*	-.396*	-.544**	-.339	-.416*	.630***	.777***	.735***	-
11	Mixed	-.090	-.415*	-.482**	-.575**	-.367*	-.333	.459*	.651***	.738***	.783***

*Note: p < .05 *, p < .01 ** and p < .001 ***.*

Table 9
Pearson correlation between arithmetic performance (TTR), symbolic and non-symbolic RAN for both age groups.

16-30 years		M	SD	1	2
1	Symbolic RAN tasks	14.35	2.05	-	
2	Non-symbolic RAN tasks	18.50	2.32	.800***	-
3	TTR tasks	29.06	5.60	-.745***	-.638***
60-85 years					
1	Symbolic RAN tasks	17.90	3.89	-	
2	Non-symbolic RAN tasks	24.82	3.04	.741***	-
3	TTR tasks	26.33	5.13	-.501**	-.492**

*Note: p < .05 *, p < .01 ** and p < .001 ***.*

Another Pearson correlation was conducted to investigate if there was a higher correlation between arithmetic and the symbolic RAN (digit, roman numbers) or arithmetic and the non-symbolic RAN performance (canonical, non-canonical, dice, dots). Symbolic and non-symbolic RAN performances revealed negative and high correlations with arithmetic performance (r = -.60 and r = -.55, p = .00, respectively).

Here we went in further detail and conducted a Pearson correlation for the above-mentioned measure, to identify the difference in both age groups. Table 9 showed that there are high and negative significant correlations between the symbolic and the non-symbolic RAN in regard to the arithmetic tasks for both age groups.

In addition, a Spearman correlation has been conducted to analyze the correlation between the performance of the participants in the arithmetic tasks and their frequency of solving riddles. The Spearman correlation coefficient revealed that there is no significant correlation between the frequency of solving riddles and the performance in the arithmetic tasks (p = .79). In this case we cannot interpret this result any further.

Several Spearman correlations have been conducted to investigate the correlations between RAN performance and the frequency of reading and of solving riddles. In regard to the frequency of reading the Spearman's rho correlation coefficient revealed that there was a negative moderate correlation between RAN performance and the participants frequency of reading, (ρ = -.49, p = .00). We can state that the more the participants read, the faster they were in the RAN tasks. Regarding the frequency of solving riddles the Spearman's rho correlation coefficient revealed that there was also a negative moderate correlation between RAN performance and the frequency of the participants solving riddles, (ρ = -.32, p = .00). This means that participants who frequently solved riddles, showed better RAN performance.

4. Discussion

The main goal of our research was to focus on differences between younger and older adults in RAN and arithmetic performance. In detail, we hypothesized younger adults to be faster in their RAN performance and to solve more addition and subtraction tasks than older adults, as these rely to a larger degree on executive functions that degrade with age than multiplication and division tasks. Therefore, we compared the number of correctly solved tasks within the set time of one minute. Furthermore, we hypothesized a faster processing time in RAN tasks to correlate with a higher number of correctly solved arithmetic items in both age groups. Lastly, we hypothesized the relation between symbolic RAN and arithmetic to be higher than between non-symbolic RAN and arithmetic in adults. As a general result, all our hypotheses could be confirmed by our findings.

In our sample, younger adults showed significantly faster processing times in all types of RAN compared to the older age group. Given that RAN is a measure of processing speed, the slower results of older adults are in line with the age-related decrease of the latter.

Both age groups showed the highest performances in digit RAN, which is hardly surprising considering the fact that digits are the most common form of representation for numeric information. The slowest responding times were measured in the non-canonical finger-configuration task in both age groups as well. This newly introduced task presents an unusual and not common finger-configuration which makes serial naming less automatized (Hornung et al., 2017).

We were also able to prove our assumption in relation to the arithmetic performances. Younger adults only solved more arithmetic problems in the addition and subtraction tasks. Older adults did not solve significantly fewer division and multiplication tasks. This might be drawn upon the fact that the retrieval of the answers to single-digit tasks is not affected by aging effects. More specifically, the

single-digit multiplication tasks are extensively studied from school age as part of multiplication tables. Multiplication uses more retrieval than addition or subtraction, which might explain our findings. Although division requires less retrieval than multiplication (Imbo & Vandierendonck, 2007), we assume that older people can still make use of retrieval for division tasks due to a longer experience with arithmetic (Hinault & Lemaire, 2016). However, the exact reason for the lack of differences between the age groups in the division tasks is unknown and thus this only provides a potential explanation. Furthermore, there was a significant negative correlation between RAN reaction times and arithmetic performance. It can be stated that the predictor function of RAN which has been observed in children may also be found in adults across different age groups. In younger adults the correlations were more pronounced than those in the older age group. Previous studies usually relied on color and object RAN when measuring non symbolic RAN. More specifically, in children the relationship between arithmetic and RAN was especially high between arithmetic and non-symbolic RAN relying on numeric stimuli such as finger configurations and dice pattern. Therefore, we administered non-symbolic RAN tasks that involved numeric stimuli such as dice, canonical and non-canonical finger configurations and dots, in order to test whether and how these correlations may change over the lifespan. As suggested in the introduction, our results show that symbolic RAN (digit and roman number RAN) correlated the highest with arithmetic in both age groups due to a long-term practice with symbolic stimuli like numbers. Concerning the roman numbers, we cannot assume that the amount of practice of these stimuli is comparable to the practice of digits. Nevertheless, our results showed similar correlations for roman numbers as they did for digits. This relation appears to be different when the RAN tasks are considered individually. Our results revealed that the non-symbolic dots-RAN shows the highest correlation with arithmetic over the total sample. Since the dots-RAN requires a small calculation operation rather than an automatized retrieval, it seems that this RAN task has a good predictor ability for arithmetic fluency.

When we analyzed the correlation between the different RAN and arithmetic tasks, we found interesting results. In the younger age group, the correlation between roman numbers and addition was the highest. As mentioned before, simple addition tasks use up to 88% retrieval (Imbo & Vandierendonck, 2007), which might support our findings. Contrary to our expectations, the highest specific correlation in the older age group was found between dots-RAN and multiplication. This seems to be less obvious, since the dots tasks are rather similar to a small addition operation. While executing the dots-RAN, the single dots are grouped into larger units and then added together, which does not resemble a multiplication operation. This aspect needs to be analyzed more extensively to detect possible reasons for this relation.

All in all, the newly introduced RAN-tasks (roman numbers, dots and non-canonical fingers) appeared to be a good complement to the existing ones. However, the non-canonical RAN only showed significant correlation with arithmetic in the younger age group. It was also the task with the slowest reaction times regardless of the participants´ age. Thus, this task has to be reviewed further in regard of its possible applications.

The questionnaire was used to enquire further information, such as demographic and socioeconomic data. We also wanted to collect information about leisure activities, like the frequency of reading and solving riddles and if there is a relation to the RAN tasks. The significant correlation between the frequency of reading and the RAN performance is obvious since this correlation has already been shown several times. It is positive that the correlation also persists for numerical RAN tasks, which supports their good functionality. Regarding other leisure activities, no significant correlations could be found, so we did not investigate these analyses further.

5. Limitations and Outlook

We investigated the differences in RAN and arithmetic tasks in two different age groups. One limitation of the present study is that the sample size was relatively small for the younger (*N* = 33) and the older age group (*N* = 31). A larger sample size would strongly support our findings in general. Nevertheless, our study can lay a valuable foundation for further research. As a second limitation could be mentioned that a standardisation was not totally fulfilled because the external conditions were not the same for the participants as the study was conducted at the participants' home. Furthermore, the administration of the tasks could have minimal deviations as the time was stopped by hand. Nevertheless, it was determined that the time was only stopped from the mention of the first symbol in the RAN and the first response in the arithmetic tasks. Thereby we tried to standardise the tasks for all participants at least. A better way to collect data would be to administrate the tasks under laboratory conditions. In addition, it has to be noted that the data of the questionnaire is based on the participants' own statements, thus subjective. Therefore, one has to be careful in interpreting the results. Most importantly, we suggest for further studies to distinguish between single- and multi-digit tasks in arithmetic as they require different cognitive processes. The multi-digit calculations rely less on retrieval and thus age-related effects could be expected in all types of arithmetic operations. All in all, this study provides valuable findings which can lay a foundation for further research within the domain of RAN and arithmetic throughout adulthood.

Appendix 1

Table 3
One-way MANOVA for specific arithmetic performances for both age groups.

	16-30 years (*N* = 33)		60-85 years (*N* =31)				
	M	SD	M	SD	F	p	η^2
Addition	33.27	3.65	30.65	3.80	7.96	<.01	.114
Subtraction	30.39	4.98	27.03	5.19	7.01	<.05	.102
Multiplication	26.30	6.98	26.00	6.01	0.03	.853	.001
Division	26.88	8.07	22.77	8.83	3.77	.057	.057
Mixed	28.45	6.10	25.19	5.46	5.05	<.05	.075

References

APA Dictionary of Psychology. (2014). Phonological Awareness. In *APA Dictionary of Psychology*. Retrieved January 11, 2022, from https://dictionary.apa.org/phonological-awareness

De Smedt, B., & Boets, B. (2010). Phonological processing and arithmetic fact retrieval: Evidence from developmental dyslexia. *Neuropsychologia*, *48*(14), 3973–3981. https://doi.org/10.1016/j.neuropsychologia.2010.10.018

Denckla, M. B., & Cutting, L. E. (1999). History and significance of rapid automatized naming. *Annals of Dyslexia*, *49*(1), 29. https://doi.org/10.1007/s11881-999-0018-9

Geschwind, N., & Fusillo, M. (1966). Color-Naming Defects in Association With Alexia. *Archives of Neurology*, *15*(2), 137–146. https://doi.org/10.1001/archneur.1966.00470140027004

Gordon, P. C., Islam, A. T., & Wright, H. H. (2021). Rapid automatized naming (RAN): Effects of aging on a predictor of reading skill. *Aging, Neuropsychology, and Cognition*, *28*(4), 632–644. https://doi.org/10.1080/13825585.2020.1806987

Hinault, T., & Lemaire, P. (2016). Chapter 10 - Age-related changes in strategic variations during arithmetic problem solving: The role of executive control. In M. Cappelletti & W. Fias (Eds.), *Progress in Brain Research* (Vol. 227, pp. 257–276). Elsevier. https://doi.org/10.1016/bs.pbr.2016.03.009

Hornung, C., Martin, R., & Fayol, M. (2017). General and Specific Contributions of RAN to Reading and Arithmetic Fluency in First Graders: A Longitudinal Latent Variable Approach. *Frontiers in Psychology*, *8*, 1746. https://doi.org/10.3389/fpsyg.2017.01746

Imbo, I., & Vandierendonck, A. (2007). Do multiplication and division strategies rely on executive and phonological working memory resources? *Memory & Cognition*, *35*(7), 1759–1771. https://doi.org/10.3758/BF03193508

Jacobson, J. M., Nielsen, N. P., Minthon, L., Warkentin, S., & Wiig, E. H. (2004). Multiple Rapid Automatic Naming Measures of Cognition: Normal Performance and Effects of Aging. *Perceptual and Motor Skills*, *98*(3), 739–753. https://doi.org/10.2466/pms.98.3.739-753

Koponen, T., Georgiou, G., Salmi, P., Leskinen, M., & Aro, M. (2017). A meta-analysis of the relation between RAN and mathematics. *Journal of Educational Psychology*, *109*(7), 977–992. https://doi.org/10.1037/edu0000182

LeFevre, J. , Sadesky, G. S. & Bisanz, J. (1996). Selection of Procedures in Mental Addition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22* (1), 216-230. https://doi.org/10.1037/0278-7393.22.1.216

Lervåg, A., & Hulme, C. (2009). Rapid Automatized Naming (RAN) Taps a Mechanism That Places Constraints on the Development of Early Reading Fluency. *Psychological Science*, *20*(8), 1040–1048.

Moeller, K., Klein, E., Fischer, M. H., Nuerk, H.-C., & Willmes, K. (2011). Representation of Multiplication Facts-Evidence for partial verbal coding. *Behavioral and Brain Functions*, *7*(1), 25. https://doi.org/10.1186/1744-9081-7-25

Radvansky, G. A., Zacks, R. T., & Hasher, L. (1996). Fact retrieval in younger and older adults: The role of mental models. *Psychology and Aging*, *11*(2), 258–271. https://doi.org/10.1037/0882-7974.11.2.258

Thevenot, C., Fanget, M., & Michel, F. (2007). Retrieval or nonretrieval strategies in mental arithmetic? An operand recognition paradigm. *Memory & Cognition*, *35*, 1344–1352. https://doi.org/10.3758/BF03193606

Wiig, E. H., Nielsen, N. P., Minthon, L., McPeek, D., Said, K., & Warkentin, S. (2002). Parietal Lobe Activation in Rapid, Automatized Naming by Adults.

Perceptual and Motor Skills, *94*(3_suppl), 1230–1244. https://doi.org/10.2466/pms.2002.94.3c.1230

The brain on numbers - The relation between numerical formats probed with electroencephalography

Zuzana Brandt, Leon Erich Geibel, Jonna Krier, Anton Franz Lachmann, Aurélie Marochi, Franziska Wagner

Supervision: Dr. Mila Marinova

The integration of numerical formats and discrimination between them in the brain are a highly debated topic in the field of neurocognitive research. It is assumed that adults have two different number processing systems, a symbolic one, which processes symbols, for instance letters or Arabic digits and a non-symbolic one. Previous studies in this field showed that adults, depending on the notation, are capable of integrating the different numerical formats. However, the question whether kids are also capable of unintentional spontaneous cross-format numerical integration has not been systematically investigated yet. Therefore, the study's participants (n = 8) are German-speaking kids aged between 7 and 12. A cross-sectional design with frequency-tagged EEG technique is used. An oddball design with numbers smaller than five as standard stimuli and numbers larger than five as deviant stimuli was applied. Participants' neural activity in 12 different conditions, 3 single notation trials (digits, words, dots) and 3 mixed notation trials (digits-dots, words-dots, words-digits) was measured. For each notation there was an experimental and a control condition. Significant oddball responses were observed in the single notation trials, supporting the hypothesis of an unintentional integration between small and large numbers in children. Furthermore, as significant responses were also found in the mixed notation trials, implying kids are able of integrating numeral information across different numerical formats, the second hypothesis is also supported and in line with previous findings in adults (Marinova et al., 2021).

Key words: children, frequency-tagged EEG, numerical integration, oddball design

Introduction

Numerically literate people have at least three ways of representing numbers: as Arabic numerals (e.g."3"), as number words (e.g. "three"), and as a collection of items such as dot arrays (e.g. "●●●"). The relation between different numerical formats and their underlying brain mechanisms is a highly discussed topic in psychological research. So far, most studies on this topic have been conducted on adults. Findings suggest that adults can relate and integrate different number notations (digits, dots, number words). How young children process the different numerical formats, however, has not been systematically investigated yet. Therefore, in the current study we aim to investigate the neuro-cognitive underpinnings of numerical integration in children aged between 5 and 12.

Traditional research on numerical cognition has shown that there exist two main systems that children and adults use to connect non-symbolic and symbolic numerical stimuli and symbols. These two systems are the Approximate Number System (ANS) and the Object Tracking System (OTS), also known as the parallel individuation system (Reynvoet & Sasanguie, 2016). To start with, the Approximate Number System refers to a non-verbal and analogue representation of numbers which means that the non-symbolic representation of numbers is not precise. Other than that, in the Object

Tracking System the representation of numbers is precise but there is a limited capacity of this system (up to three or four). Because the ANS has an unlimited capacity for representing numbers, it is called the main center for the representation and correlation of numbers and led the researchers to the ANS theory.

The ANS theory says that the non-symbolic processing of numbers is inborn. The non-symbolic representation of the number is processed by an automatic valuation of the quantity of the numbers represented. The theory asserts that physical characteristics do not have an impact on the processing of numerical stimuli. A modern alternative to this theory is the ANS mapping account. Following this, researchers assume that the processing of symbolic numbers is closely linked to the processing of the matching non-symbolic numbers. This means that for adults the numerosity (i.e., the number of stimuli in a set) is automatically assigned to the appropriate number (van Hoogmoed et al., 2021)

Another study, also conducted by Marinova et al. in 2021, tested the automatic integration of numerical formats in students aged between 18 and 25. The relation between symbolic and non-symbolic numbers is located in an area in the parietal brain along the Intra Parietal Sulcus (IPS). In this study, an EEG (electroencephalogram) was used to record brain responses from participants in different tasks. The participants completed different tasks. To investigate the automatic integration of numerical magnitude and the relation between numerical formats, there were single notation trials (only number words, digits, or dots) and mixed notation trials (words-dots, words-digits and digits-dots). In all experimental trials, numbers smaller than five (1,2,3,4) were presented as standard stimuli whereas numbers larger than five (6,7,8,9) were presented as deviant stimuli. With this design in the single notation trials, it was investigated whether there is an automatic integration of small and large numbers. For the experimental condition every 5th stimulus was an oddball stimulus and for the control condition there is a randomized order of numbers presented. The stimuli were presented with a frequency of 10 Hz. The findings of this study show that there were responses recorded in the posterior scalp in the experimental conditions. The results of these study showed that there are significant brain responses for number words-digits and number words-dots but not for dots-digits. They also show that there is no automatic integration between digits and dots. Nevertheless, brain responses for the magnitude of number words and the magnitude of digits are automatically extracted. In general, the study supports the hypothesis that adults can spontaneously integrate numbers between different formats (Marinova, Georges, et al., 2021).

In our study, we focus on children and not on adults. We cannot find any studies for children on this topic yet. However, the behavior of children when relating between the different numerical formats has already been studied. While the neuro-cognitive research on numerical integration in children is rather sparse, indication of how children relate the different numerical formats, can be found in behavioral developmental literature. Previous studies have identified two models for numerical integration in children(Benoit et al., 2013; Hurst et al., 2016; Jiménez Lira et al., 2017; Marinova, Reynvoet, et al., 2021).The “quantity account” which says that children first learn how to translate between number words and dots (i.e. “three” = “●●●“) and digits and dots (i.e., “3” = “●●●“) and then later they combine these two and start to map between number words and digits (“three” = “3”).
The second account to name is the “symbolic account” which assumes that children first learn to translate between number words and dots (“three” = “ ●●●“) and number words and digits (“three” = “3”) and then later they combine these two and are able to map between digits and dots (“3” = “ ●●●“) (Marinova, Reynvoet, et al., 2021).
Even though there aren't any studies yet on children’s brain responses to numerical formats, a study was recently conducted by Marinova et al. (2021), focusing on how kindergarten children (aged between 2 and 5) relate the different numerical formats. These findings show that children experience difficulties when mapping between digits and dots. That means

that they have struggles with comparing digits and dots. It seems that they do not have problems with mapping digits and number words or number words and dots. That means that children can easily compare number words with digits or dots. Consequently, the authors interpreted this as evidence for the symbolic account. Children used the number words and their relation to digits and dots to map between digits and dots. This study is relevant for our study in order to see whether the fact that children have struggles with mapping between digits and dots is also shown in their brain responses. Besides, we are interested in whether these findings are only behavioral related or also neurological related.

Unfortunately, there is a lack of evidence in neurocognitive studies in children using EEG. Therefore, the automatic cross-format integration in children is not investigated enough. On a behavioral level, there are some results which were mentioned before. On a neurocognitive level, in fact the only proof we can extract is from the study conducted by Van Hoogmoed et al. in 2021. This study investigated, with the use of an EEG, whether the ANS theory and the ANS mapping account declare how children aged between 9 and 12 process the non-symbolic numerosity and the symbolic number. They found out that the visual characteristics of the non-symbolic stimuli are processed more automatically by the children than the count of the stimuli themselves. They also found out that activating the non-symbolic numerosity is not automatically activated by the children to process symbolic numbers (van Hoogmoed et al., 2021). Nevertheless, further research is needed to provide more details on the neurocognitive mechanisms of number processing in children. To the best of our knowledge, there are no studies yet that are systematically investigating the neurocognitive mechanisms of within and cross-format numerical processing in children. This is precisely the aim of the current study.

In our study, we investigated the brain responses in children aged between 5 and 12 (*n* = 9). Similar to the study conducted by M. Marinova et al. in 2021, the brain responses of the participants were recorded by using EEG technique.

The methods used for this study were also similar to the study conducted by M. Marinova et al. in 2021. The participants completed different tasks. There were single notation trials (only words, digits, or dots) and mixed notation trials (words-dots, words-digits and digits-dots). In all experimental trials, numbers smaller than five (1,2,3,4) were presented as standard stimuli whereas numbers larger than five (6,7,8,9) were presented as deviant stimuli. With this design in the single notation trials, we investigated whether there is an automatic integration of small and large numbers. For the experimental condition every 5th stimulus is an oddball stimulus and for the control condition there is a randomized order of numbers presented.

H1: We assume that if children automatically discriminate small vs. large numbers, we should receive significant oddball responses in the single notation trial.

H2: Furthermore, we suppose that if there was an automatic integration across numerical formats, we should obtain significant oddball responses in the mixed notation trials too.

Methods

Sample

The study was designed as a pilot study for a larger project and therefore only a small number of participants was sufficient. Our target group was primary school children aged 5 to 12 years with normal or corrected-to-normal vision. It was required that participants were schooled in German or that German was their native language. Parents were invited to accompany their children. As a compensation of participation, children received a small gadget, a book voucher of 15 euros, a picture with the cap during the EEG examination and a printed certificate of their participation.
All participants invited matched our requirements (age: *M* = 9,88; *min* = 7; *max* = 12; *SD* =

1,46) and no one was excluded due to any medical or neurological conditions or any form of learning difficulties. Recruitment yielded a total of 6 boys and 3 girls, and none of the children were excluded because of poor performance on the experimental tasks. After the data verification, we ended up with 8 correct data processing because one participant's test output showed some technical issues. In addition, all restrictions and rules related to COVID-19 were considered and were followed in the experiment.
Participants were recruited on a private basis and included our siblings, children, and their friends. The children's parents or legal guardians received an information sheet and consent form in advance and gave us their consent by signing and completing the documents. The children also gave their verbal consent. Our research project and all related documents were approved by the Ethics Committee of the University of Luxembourg before recruitment.

Task description and procedures

The children were tested in November at the premises of the University of Luxemburg in the presence of an examiner and a research assistant. During the testing phase, children performed a total of four tasks (two behavioral, one EEG, and one paper and pencil task).

TTR Task

The TTR task was used as a control measurement of mathematical abilities. Towards the end of the testing session, children were asked to complete a short math test in which they had one minute to complete as many tasks as they could on a single page without skipping any exercise. There were 5 pages with different tasks (addition, subtraction, multiplication, division and all mixed). The researcher assisted the participant and took the time. This task measured very basic math competences needed to complete the EEG task. The time needed to complete this exercise was about 5 minutes.

Reading ability task

This task aims to ensure that the children could read the number words presented at a fast rate, similar to the frequency presentation in the EEG task. The stimuli for this task consisted of German number words from one to nine. Each trial started with a 500ms fixation cross, followed by a number word presented on the screen for 166ms. After that, an inter-trial interval of 1500ms followed and then the next trial started. The participants' task was to read the word aloud as soon as they saw it. The examiner noted down whether the child responded correctly. There were 27 trials (nine number words, each presented three times). There were no training trials. PsychoPy software was used to present the stimuli. This task took approximately 5 minutes to complete. The participant sat at a distance of about 1m from the screen.

Number matching (behavioral) task

Participants had to judge whether two numbers displayed simultaneously on the screen were same (e.g., "2" "two") or different regarding their magnitude (e.g., "2" "four"). The stimuli were numbers from 1 to 9 presented as digits, number words, or dots in three cross-format conditions: 1) digits and words, 2) digits and dots, and 3) dots and words. Each trial began with a 500ms fixation cross in the center of the screen, followed by two numbers appearing simultaneously on the screen for 1500ms. Children had to answer whether the two numbers were the same or different by pressing a yellow (right arrow) or a blue (left arrow) key on a QWERTU keyboard. Responses were collected during the presentation of the stimuli up to 2000ms. Then, the next trials started after a 1500ms inter-trial interval. Prior to testing, children received 10 training trials with immediate feedback on the correctness of their performance. These were followed by 24 experimental trials without any feedback. The cross-format combinations were presented in three blocks of 24 trials in three different combinations. Written instructions were displayed on the screen prior to the trial and were also explained by the researchers to ensure they were understood

correctly. The task took approximately 5 minutes to complete.

EEG task

The EEG data were acquired at 512 Hz using 64 channels, positioned according to the integrational 10-20 system(van Hoogmoed et al., 2021). In addition, two eye electrodes were attached to the participant's left eye. Data from these electrodes were not analyzed and data was not corrected for eyeblinks. The electrode offset was held below 40 microvolts. During the actual task, participants were seated comfortably at an approximate distance of 1m from the computer screen. Children were asked to look at the computer screen while sequences of numbers in different formats in six combinations appeared: 1) only digits, 2) only dots, 3) only words, 4) digits and words, 5) digits and dots and 6) words and dots. Sequences were presented at a standard rate of 6Hz (6 stimuli per second), and each 5th stimulus (i.e. at 1.2 Hz) was a deviant stimulus. Each combination had an experimental and a control condition. The experimental condition contained oddball stimuli. The standard stimuli were always numbers smaller than five and the oddball stimuli were numbers larger than five (see figure). For the control condition, the presentation of numbers was random. Each task sequence lasted 44s, including 2s of gradual fade in, 40s of stimulation, and 2s of gradual fade out phase. The fade in and fade out phases were excluded from analysis. Each sequence was presented two times resulting in a total of 24 trials. The sequences were presented in a randomized blocked design in which all trials of the same notation were presented in one block (e.g., two trials digits experimental, followed by the two control trials). However, the presentation of the block and the order of the experimental and control conditions was counterbalanced across participants in a Latin square design. To ensure that children constantly focused on the screen, we asked them to complete a simple task. A small blue square was displayed in the center of the screen, and children were asked to press the space bar each time the color of the square changed to red. The square changed its color eight times at random intervals during the sequence. Overall, the children correctly recognized the colour change 90% of the time. The completion of this task took around 25 minutes. Due to the high concentration and possible eye fatigue, the children were allowed to take a break or drink water. After the last sequence, the cap with the electrodes was removed and the children were asked to wash, shampoo and dry their hair.

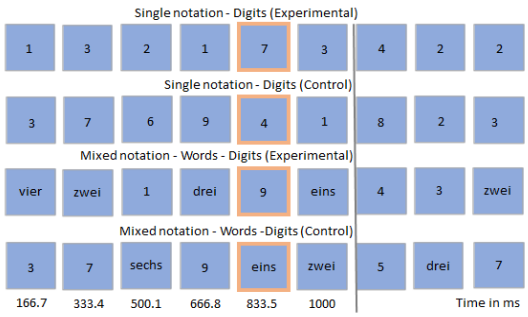


Figure 1: Visual example of experimental and control trials presented during the EEG task

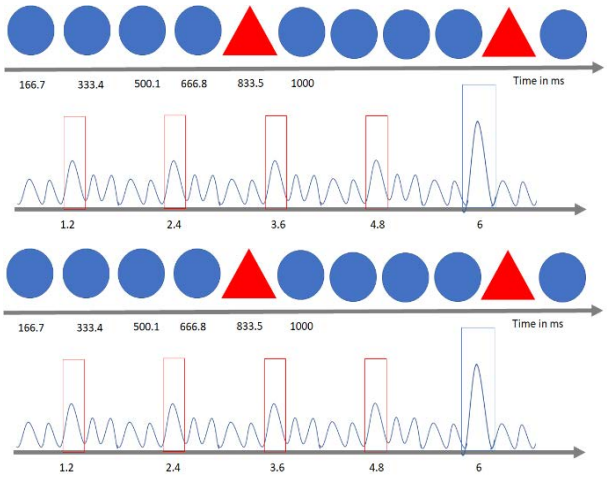


Figure 2: Depiction of paradigm and expected brain responses below

Results

Performances in the preliminary tasks and EEG data of *n* = 8 (corrected) children were analyzed.

In the TTR task, participants scored an average of *M* = 90 of 200 points (*SD* = 33.34). A

significant positive correlation with a large effect between participants' age and their TTR scores was found (r = 0.74, p = 0.035). Not surprisingly, this indicates that older children tend to perform better than younger children.

Results of the reading ability *task* show that all children responded with an accuracy of 100% (M = 1, SD = 0). None of the children made any mistake. This means that all children were capable of reading the shortly presented words (166ms-like in the EEG task) fast enough.

In the number matching task, children responded with an accuracy of 74%. They performed best in the words and digits (M = 0.82, SD = 14) and worst in the dots and digits trials (M = 0.68, SD = 0.24). The mean in the words and dots trials equals M=0.71 (SD = 0.10). These findings would indicate that this task was easiest for children when only symbolic notations were used. When symbolic and non-symbolic formats are mixed, children perform worse on average. A repeated measures ANOVA[1], however, shows that there is no significant main effect of notation (see values below).

Accuracy scores of the number matching task were submitted to a 2x3 repeated measures ANOVA with notation (digits-dots vs words-dots vs words-digits) and trial type (same vs different) as within subject factors. In the repeated measures ANOVA no main effect of notation was found, $F(2,14) = 3.03$, $p = 0.08$, $\eta_p^2 = 0.3$. Neither was a main effect of Trial Type found, $F(1,7) = 1.94$, $p = 0.21$, $\eta_p^2 = 0.22$. However, a marginally significant interaction between notation and trial type was found, $F(2,14) = 3.78$, $p = 0.049$, $\eta_p^2 = 0.35$. In a post-hoc comparison test with Bonferroni correction no significant p-values were found. This only marginally significant effect indicates that the effect of notation on amplitudes recorded is influences by the trial type, so whether the numbers' magnitude was the same or different.

All in all, the three preliminary tasks were completed sufficiently by all eight children. Their reading and arithmetic abilities meet the requirements for the completion of the EEG task.

EEG task

The EEG data was analyzed using *Letswave 6.* For all further analyses we used the baseline-corrected amplitudes.

In figure 3, four exemplary oddball topographies from our study are depicted. These topographies show neural activity at the deviant frequency meaning the oddball stimulation rate (1.2Hz) and its harmonics (2.4Hz, 3.6Hz, 4.8Hz). The red coloring indicates pronounced amplitudes in the signal at the oddball stimulation rate. The amplitudes on which the coloring of the different regions is based are baseline-corrected which means that the event related potentials (manipulation through oddballs) are subtracted from the baseline activity (internal events).

The two upper topographies show the single notation dots and the ones underneath the mixed notation words and dots. Experimental conditions are displayed on the left and control conditions on the right. Red coloring can be seen in the experimental condition only and is concentrated in the back of the brain. Even though the EEG-technique does not allow to exactly allocate neural activity to specific brain areas, it can help to define regions of interest. With the help of the topographies at hand, we grouped electrodes of the posterior scalp and defined medial-occipital (O1, Iz, Oz, O2), left occipito-parietal (P5, P7, P9, PO7) and right occipito-parietal (P6, P8, P10, PO8) as our three regions of interest. We will focus on these regions in our further analysis of neural activity.

The spectrum graph (figure 4) shows recordings from the single notation dots experimental condition in the medial occipital region. In all conditions, experimental and control, we observed a clear response at the standard frequency of 6Hz indicating that the stimulation worked and that children looked at the screen.

[1]Despite the small sample, an ANOVA was preferred over non-parametric tests for training purposes. This must also be considered for the tests still to follow when analyzing the EEG data.

Furthermore, we also observed smaller but clear peaks at the oddball stimulation rate of 1,2Hz and its harmonics (2.4Hz, 3.6Hz, 4.8Hz) in the experimental condition. There were no peaks at this frequency observed in the control condition. This indicates that our oddball design worked and thus, children's neural responses were higher when they were presented the deviant stimuli. However, to find out whether these higher amplitudes in the signal were actually significant, we performed a repeated measures ANOVA.

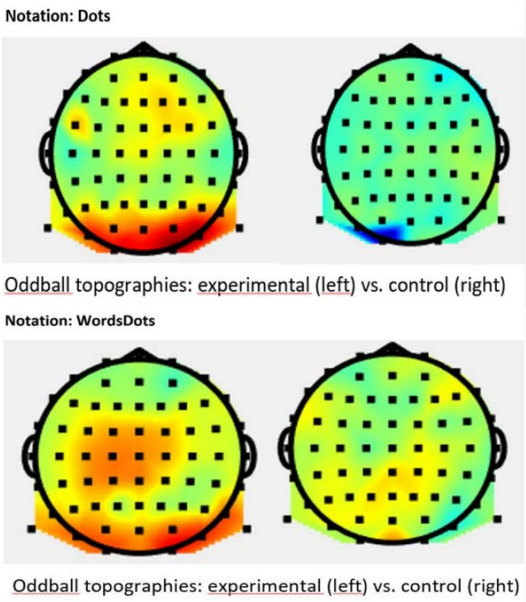


Figure 3: oddball topographies: sum of baseline-corrected amplitudes $n = 8$

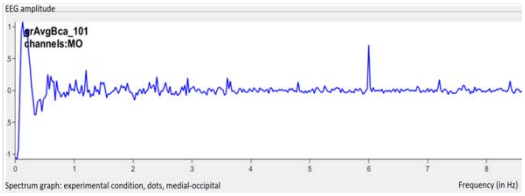


Figure 4: spectrum graph of experimental condition dots in the medial-occipital brain region

Single notation conditions

The obtained EEG data was baseline corrected in order to analyze the oddball responses. A 2x3x3 repeated measures ANOVA was conducted to compare means for the single notation conditions. The within-subject-factors were condition (experimental vs control), notation (digits vs words vs dots) and region of interest (left occipito-parietal vs right occipito-parietal vs medial occipital). Whenever Mauchly's test of sphericity indicated that the assumption of sphericity is violated, the Greenhouse-Geisser correction was applied.

Main effects

There was a main effect of condition in the single notation, $F(1,7) = 41.33$, $p < .001$, $\eta_p^2 = 0.86$ with experimental conditions yielding significantly stronger amplitudes than control conditions, $p_{bonf} < .001$. The comparisons of experimental conditions to control conditions are

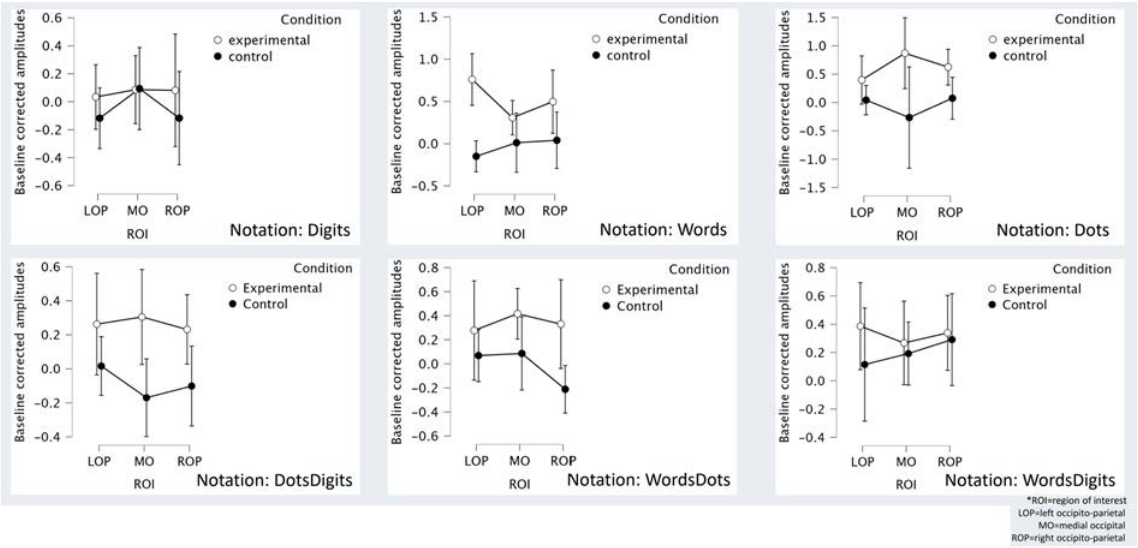


Figure 5: JASP-output, graphic representation of 2x3x3 repeated measures ANOVA

depicted in figure 5. There was no significant main effect of notation, $F(1.16,8.14) = 4.64$, $pGG = 0.06$, , $\eta_p^2 = 0.40$. There was no significant main effect of region of interest either, $F(2,14) = 0.25$, $p = 0.78$, , $\eta_p^2 = 0.03$.

First-order interactions

There was no significant notation x condition interaction, $F(2,14) = 2.04$, $p = 0.17$, , $\eta_p^2 = 0.226$. Furthermore neither the notation x region of interest interaction $F(4,28) = 0.91$, $p = 0.47$, , $\eta_p^2 = 0.12$ nor the condition x region of interest interaction $F(2,14) = 0.09$, $p = 0.92$, $\eta_p^2 = 0.01$ was significant

Second-order interaction

The three way interaction notation x condition x region of interest was not significant $F(4,28) = 2.37$, $p = 0.08$, $\eta_p^2 = 0.25$.

Mixed notation conditions

The obtained EEG data was baseline corrected. A 2x3x3 repeated measures ANOVA was conducted to compare means for the mixed notation conditions similarly to the single notation conditions. The within-subject-factors were condition (2 levels), mixed notation (3 levels) and region of interest (3 levels).

Main effects

Similar to the single notation, there was a main effect of condition in the mixed notation, $F(1,7) = 9.71$, $p < .05$, $\eta_p^2 = 0.58$ with experimental conditions yielding significantly stronger results than control conditions, $p_{bonf} < .05$. There was no significant main effect of notation, $F(2,14) = 1.57$, $p = 0.24$, $\eta_p^2 = 0.18$. There was also no significant main effect of region of interest, $F(2,14) = 0.26$, $p = 0.78$, $\eta_p^2 = 0.04$.

First-order interactions

There was no significant notation x condition interaction, $F(1.22, 8.57) = 1.56$, $pGG = 0.25$, $\eta_p^2 = 0.18$. The notation x region of interest interaction $F(4,28) = 1.52$, $p = 0.22$, $\eta_p^2 = 0.18$ was not significant. The condition x region of interest interaction was also not significant, $F(1.19, 8.34) = 0.11$, $pGG = 0.80$, $\eta_p^2 = 0.02$.

Second-order interaction

The interaction notation x condition x region of interest was not significant, $F(4,28) = 1.85$, $p = 0.15$, $\eta_p^2 = 0.21$.

Discussion

This study was designed to examine first, whether children can spontaneously and unintentionally discriminate between small and large numbers within a numerical format (e.g., digits, words, dots), and second if there is an automatic integration across numerical formats (i.e., digits – words, digits – dots, dots – words). If this was the case, then significant oddball responses in the mixed notation trials should be present. These hypotheses were tested by the EEG frequency-tagged technique.

The results for the single notation trials obtained by a repeated measures ANOVA showed that there is a main effect of condition indicating that significant oddball responses were recorded in the experimental trial. This means that when comparing the experimental with the control condition, when participants were presented the deviant stimulus, they (>5) had higher neural activity across the posterior scalp which leads to the assumption that there is an automatic discrimination between small and large numbers in children. This result is obtained in every single notation control condition, no matter of the type of notation. It is possible to infer that as early as seven years old, children are capable of processing the semantic of numbers rapidly. However, the results did not show any significant interaction between notation versus condition, conditions versus region of interest, and notation versus region of interest.

On the same line, the results obtained for the mixed notation trials also show that children are

capable of integrating the semantic numerical information within different numerical formats (i.e., word numbers, dots and Arabic numerals), as significant oddball responses were found in the digits–words and words–dots conditions. These findings lead to the conclusions that the experimental condition also showed higher neural activity when compared to the control condition in the mixed notation trials when presented the deviant stimulus (>5). Concluding, as a result, there is evidence to support that children are capable of integrating different numerical formats across numerical formats. Regarding the effect of notation and region of interest, there was no significant main effect.

Even though it was not statistically significant, lateralization of the results was found in this study on the experimental condition. First, analyzing the results for dots, pronounced amplitudes were found in the medial occipital region of the brain with the oddball stimulation. Second, when presented word numbers, the amplitudes were higher on the left occipital region of the brain. And finally, with digit numbers the lateralization of the results was placed in the right occipital region.

In addition, the results obtained in this study with children are in line with previous studies with adults, for instance in Marinova et al., 2021. In this previous literature the results pointed a significant magnitude-related oddball response within all the experimental conditions, however, no response in the control condition. This means that all participants in the study (Marinova, Georges, et al., 2021) could differentiate the different number notations. On the hand, their results on mixed notation trials showed significant oddball responses in some crossed notation trials as, digit-words and word-dots, but not with dots-digits. They also concluded that there is evidence that digits-words and words-dots are easily integrated, forming an abstract representation.

A limitation of this study is the small sample size of eight participants only. In the field of neuroscience research, low statistical power (brought by restricted data) is usually a common problem. This condition reduces the probability that a statistically significant result is trustable. A study in this field was conducted (Button et al., 2013), showing that the average statistical power of studies in the neurosciences is very low. Considering that, some consequences come along with a small sample size, such as overestimating the effect size and low reproducibility of the results. In further studies a vaster sample should be used for the analysis, making it possible to draw valid and robust conclusions in this research area. However, despite the limitations, this study is the first one in this research area showing rapid and unintentional processing of single and mixed format trials in children by using the Frequency-tagged EEG methodology. Additionally, it is possible to exclude unintended effects on the EEG data due to serious reading or arithmetic difficulties of children, since all participants performed sufficiently high in the preliminary tasks.

In sum, the findings in this study, show that children could automatically and spontaneously discriminate small and large numbers, and also integrate different number notations that were presented to them. Results support our two hypotheses that were settled in the beginning of the research.

References

Benoit, L., Lehalle, H., Molina, M., Tijus, C., & Jouen, F. (2013). Young children's mapping between arrays, number words, and digits. *Cognition*, *129*(1), 95–101. https://doi.org/10.1016/j.cognition.2013.06.005

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

Hurst, M., Anderson, U., & Cordes, S. (2016). Mapping Among Number Words, Numerals, and Non-Symbolic Quantities in Preschoolers. *Journal of Cognition and Development*, *18*. https://doi.org/10.1080/15248372.2016.1228653

Jiménez Lira, C., Carver, M., Douglas, H., & LeFevre, J.-A. (2017). The integration of symbolic and non-symbolic representations of exact quantity in preschool children. *Cognition*, *166*, 382–397. https://doi.org/10.1016/j.cognition.2017.05.033

Marinova, M., Georges, C., Guillaume, M., Reynvoet, B., Schiltz, C., & Van Rinsveld, A. (2021). Automatic integration of numerical formats examined with frequency-tagged EEG. *Scientific Reports*, *11*(1), 21405. https://doi.org/10.1038/s41598-021-00738-0

Marinova, M., Reynvoet, B., & Sasanguie, D. (2021). Mapping between number notations in kindergarten and the role of home numeracy. *Cognitive Development*, *57*, 101002. https://doi.org/10.1016/j.cogdev.2020.101002

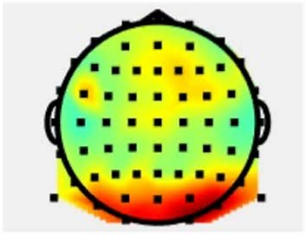
Reynvoet, B., & Sasanguie, D. (2016). The Symbol Grounding Problem Revisited: A Thorough Evaluation of the ANS Mapping Account and the Proposal of an Alternative Account Based on Symbol–Symbol Associations. *Frontiers in Psychology*, *7*, 1581. https://doi.org/10.3389/fpsyg.2016.01581

van Hoogmoed, A. H., Huijsmans, M. D. E., & Kroesbergen, E. H. (2021). Non-Symbolic Numerosity and Symbolic Numbers are not Processed Intuitively in Children: Evidence From an Event-Related Potential Study. *Frontiers in Education*, *6*, 241. https://doi.org/10.3389/feduc.2021.629053

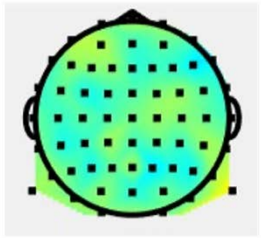
Appendix

Oddball topographies: sum of baseline-corrected amplitudes n=8

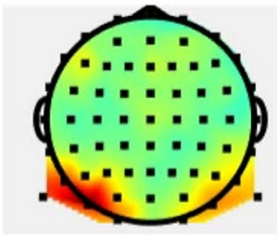
Dots - Exp



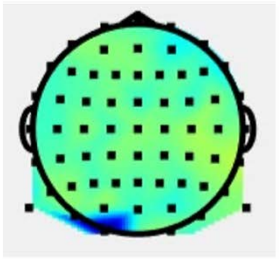
Digits – Exp



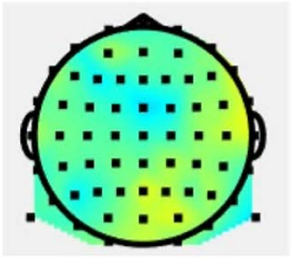
Words - Exp



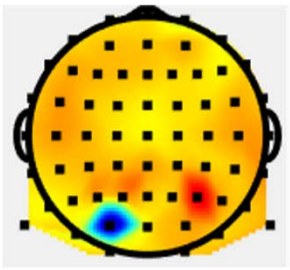
Dots – Contr



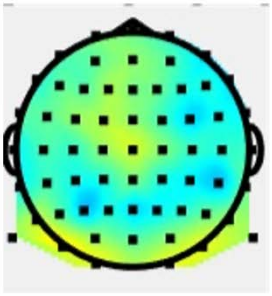
Digits – Contr



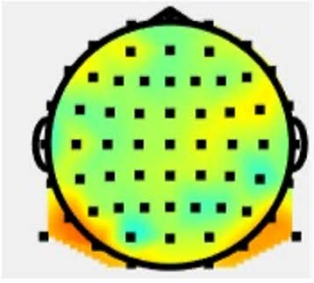
Words – Contr



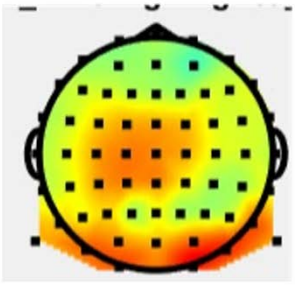
Dots-Digits Exp



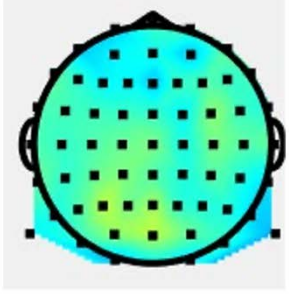
Digits-Words Exp



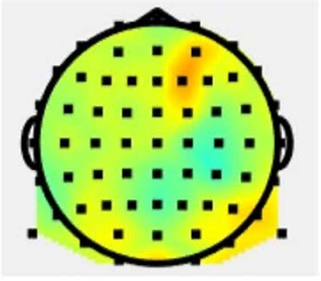
Words-Dots Exp



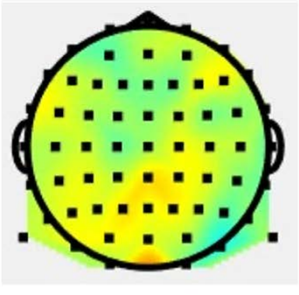
Dots-Digits Contr



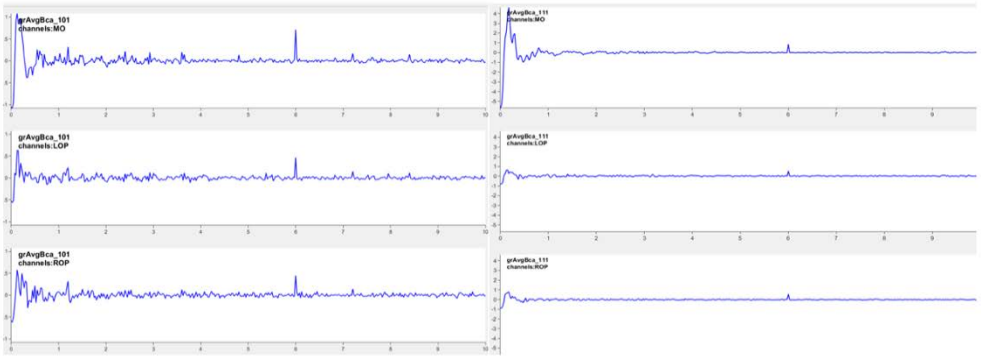
Digits-Words Contr



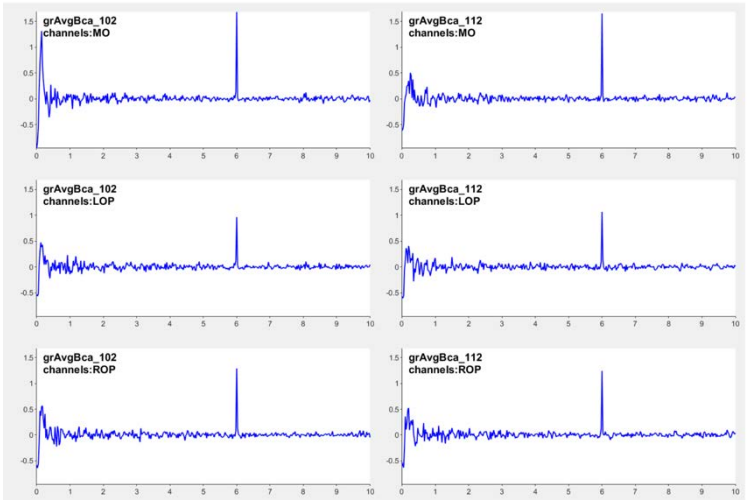
Words-Dots Contr



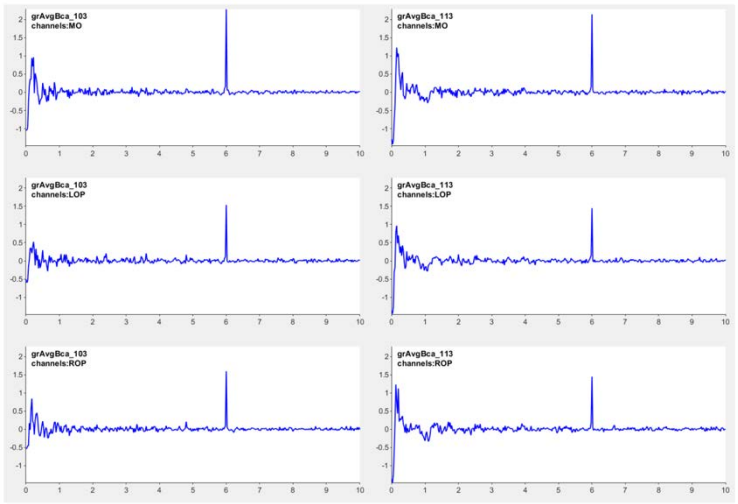
Spectrum graphs (only ROI):
Dots (Exp vs Control)



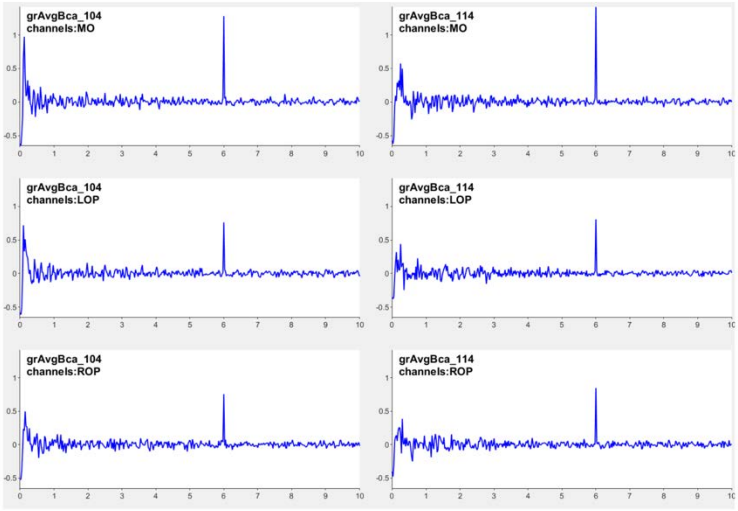
Digits (Exp vs Control)



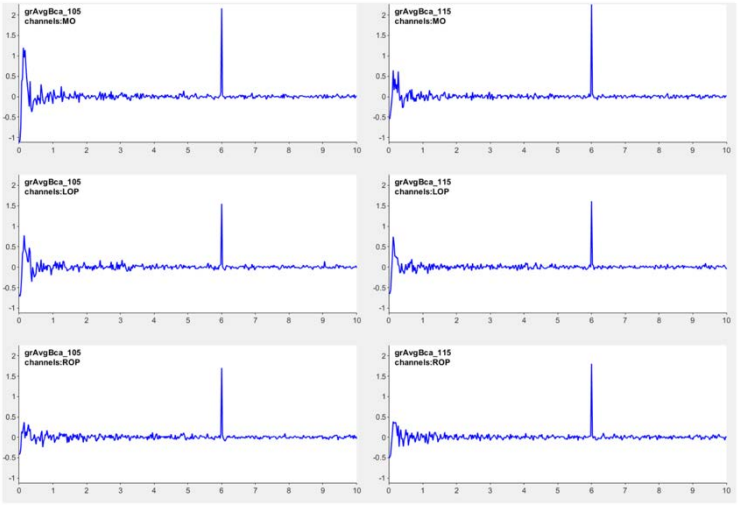
Words(Exp vs Control)



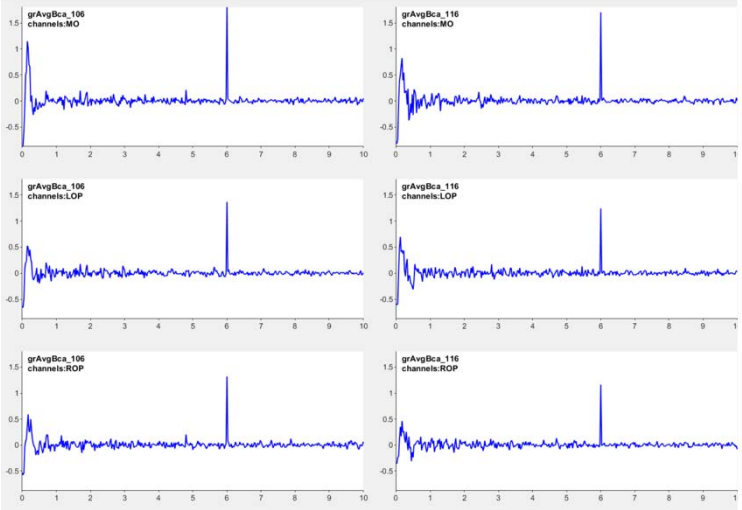
Dots – Digits (Exp vs Control)



Words – Digits (Exp vs Control)



Words – Dots (Exp vs Control)



Implicit learning of color-number associations

Andreas Bieck, Diogo Da Silva, Susanne Fuchs, Marielle Mousel, Luisa Musfeld and Melissa Pagliai

Supervision: Talia Retter

In a color-number contingency learning paradigm, we analysed the association between color and numeric concepts (parity and magnitude). Numbers are associated with colors by presenting them with a high probability in that color. Learning is indicated by reduced performance for low-probability trials in terms of accuracy (of the color assignment task) and response time. We examined if the numerical concepts of parity and magnitude are implicitly recognized and automatically associated with high probability colors. To this end, the reported experiments use five blocks of about 110 trials with 5-15 congruent (high-probability) trials followed by one incongruent (low-probability) trial. A sixth block with black double-digit numbers was used to test explicit color association recall. For parity the predicted effect on accuracy is significant but not for response time. For magnitude, the predicted effect on accuracy and response time are both significant. These findings support the thesis of the implicit learning of color-number associations at a conceptual level and set the stage for new research on color associations in an educational context.

Introduction

Texts as well as numbers can be systematically presented in different colors, but there is limited evidence on how this affects our conceptual understanding. An everyday example is the use of highlighters which is therefore a possible starting point for further investigation on that topic. By using different colors, different elements from a text can be highlighted and a conceptual meaning or assignment can be attributed to them. Thematically related elements can be marked in the same color, for instance while learning a new language all the irregular verbs in a text can be easily identified. There are many other examples that can be added to the previous one, as text-markers are mostly used to support learning scenarios. In a survey conducted in 2007, the “Statista Research Department” asked more than 10,000 people about their use of text-markers. About 57% of the participants said that they used them occasionally and 8% said that they used them frequently which is contrasting with the remaining 35% who never used such tools. The question therefore arises as to how efficient this strategy is, i.e. highlighting in a learning-related environment conceptually related content in specific colors. The extent to which learning-improving effects are present remains open. The question that arises is whether a color related effect on associative learning of general concepts presented in a text or in a sequence of numbers exists?

A study by Rinaldi & al. (2019) *"Do the colors of Educational Number Tools Improve Children's Mathematics and Numerosity?"* already focused on the topic of whether color associations concerning sequences of numbers influences children’s mathematical ability or understanding. Colored teaching materials such as “Numicon” (Oxford University Press, 2018) and “Numberjack” (Ellis, 2006) were used to check to what extent the color positively or negatively influenced the children's mathematical performance. Positive effects for Numicon could be observed with regard to the children's perception of the size (magnitude) of a number in contrast to the use of Numberjack. However, no positive effect was found for general mathematical skills. Nevertheless, the

difference in children's perception of numbers between Numicon and Numberjack is very interesting because in Numberjack random and repetitive colors are used for different numbers, whereas in Numicon the coloring of numbers is much more structured. A connection between color cognition and numerical processing seems to apply in this case. While a partly positive effect was found for numbers, unfortunately no positive influence of color associations on children's mathematical abilities could be observed. Another point that remains open is whether these effects could also be identified in a text. Colors can be well associated with text passages or number sequences, but such associations do not necessarily lead to a meaningful conceptual framework with appropriate learning outcomes.

As previously described, research has shown that, conversely, conceptual structures can also have an influence on color perception. Such a connection is principally demonstrated by the study of Athanasopoulos et al. *"Cognitive representation of colors in bilinguals: The case of Greek blues"*. The researchers found a fundamental influence of linguistic structure on color perception. Individuals who speak different languages, and thus have access to different linguistic lexicons as well as different grammar, seem to perceive their environment differently. In the study itself, the researchers compared bilingual Greek speakers with monolingual English speakers. The individuals were presented with different shades of blue. It was found that Greek-speaking individuals could more easily perceive those color nuances. In Greek, compared to English, there are several different terms that differentiate the color blue into different shades. The researchers therefore concluded that there must be a connection between the representation of specific color terms in linguistic memory and their internalization to the individual perception of the color blue. It was thus confirmed that conceptual knowledge can influence color perception.

However, the general question remains whether color associations can be used to capture higher-level conceptual understanding. Therefore, we want to address this issue in our study. We investigate potential color association effects with respect to the following two numerical concepts: number size (magnitude) and parity. The colors blue (even) and yellow (odd) are used for the concept of parity. Regarding magnitude, we use the colors red (large) and green (small). The task of the participants here is to reproduce the single color. The numerical properties are thus only indirectly recorded. The aim is to compare the impact of the concept associations in their effectiveness by means of congruent and incongruent runs. The effectiveness is not only determined by the correctness of the color assignment, but also by the speed of the response. The respective item associations can then be compared with each other on the basis of these aspects. The merely indirect, i.e. automatic, recording of the numerical properties of the colored numbers is particularly essential in our procedure. To ensure this, we use an implicit associative learning paradigm in which numbers are associated with colors by presenting them with an increased probability in that color. This paradigm is also known as the "contingency learning paradigm" (see also: "statistical learning").

In a study of Schmidt & al. (2007), *"Contingency Learning without awareness: Evidence for implicit control"*, the experimental design is also based on a contingency paradigm. Some words had a high-probability (75% to 50%) of appearing in a specific color, which represented therefore high contingency word color pairs or associations. From time to time, the word would appear in a different color. This represents low contingency word color pairs. Uncolored distracter words were also shown. To be able to measure the effect of associative learning, the reaction time of the color recognition was compared for high contingency vs. low contingency trials. Crucially, the researchers also found that this effect worked for individuals who were not aware of the contingency. The results thus indicate that the contingency effect does not necessarily depend on directed attention, but can work in an implicit way.

In a separate Study by Schmidt & al (2018), the authors investigated if the color-word contingency paradigm could also be applied to *categories of words* in contrast to the

previously described investigation on simple single item level word-color associations. Based on their experimental results they concluded that "a category based contingency effect was observed", i.e. learning was indeed influenced by contingency when applied to categories and not only when applied to items. Just as in the study already discussed earlier, the associative learning model of contingency thus also seems to be confirmed here. The results are useful to imagine new learning models and in the context of our experiment, it spreads light on potential relationships between colors and conceptual associations, thus encouraging us also to investigate new numerical topics, such as parity, magnitude and further numerical categories.

Bankiers and Aslin (2017) used a high probability associative learning paradigm to study implicit associative learning in synesthetes and non- synesthetes. Synesthetes automatically associate different senses, for example a sound can be associated with a color. Bankiers and Aslin especially investigated the association between colors and geometric "snowflake" shapes. The results showed that associative learning occurs differently for both groups, with interferences of low contingency pairings being more important for synesthetes. Although our study is not concerned with synesthesia, it is useful to understand how people can implicitly learn some color object associations.

Lin and Mc Lead's (2018) study *"The acquisition of simple associations as observed in color-word contingency learning"*, also examined the learning of word-color associations by the means of contingency effects. Three words were each assigned a different color. The reaction time was measured in relation to the response speed of the individuals studied. These results were also compared with reference values of words that were not colored. If the contingency of the word with the associated color was high, the reaction time was shorter than if the contingency was low. Thus, with the help of the effect of contingency, associative learning processes were again demonstrated. In our study, color associations formed with implicit learning will be used to probe the numerical concepts of parity, specifically we ask whether numerical concepts will be implicitly and automatically associated with high probability colors. It remains debated whether parity is accessed automatically, that is, without explicit attention to this property.

Reynovoet & al (2002) proposed in their study *"Automatic Stimulus response associations may be semantically mediated"* that parity is automatically processed. The researchers showed that the response to stimulus is mediated by parity. To confirm that, they presented numbers using tachistoscope-methods based on a visual instrument or screen that flashes a series of images onto a screen at a rapid speed to test perception memory and learning. It was underlined that the time to react is longer if the two numbers have a different parity, as well as other different properties, for example if the numbers are not close to each other, if they are not part of the same group of numbers, and if the modality is different. Automatic processing is typically agreed upon for magnitude. However, other studies suggest that parity is not automatically processed.

There are not many other studies demonstrating the automatic nature of processing, which motivated us to further investigate in such a direction. Colors and numbers may have widely been studied independently in psychological research, however, the way colors might be used to measure numerical concepts is a pretty new topic. That is why we decided to conduct some research to understand whether colored numbers can be used to measure whether people automatically process parity and magnitude, expecting that concept level numerical color associations will be present for parity and magnitude.

Methodology

Participants

34 people participated in the study in exchange for €10 in vouchers and, if requested, a signed participation hour as compensation. Participants' age ranged from 19 to 27 years old; their mean age was 21,68 years old. There were 27 female, 6 male and 1 non-binary

participants; 29 of whom reported being right-handed, 5 left-handed and none ambidextrous. All reported normal or corrected-to-normal visual acuity, the absence of any learning disabilities (like dyscalculia), and that they did not experience synesthesia. Their first language of math education was German (29 subjects), French (4 subjects) or other (1 subject).

The participants were recruited by a recruitment flyer through an online platform of the University of Luxembourg. The testing took place from November 2021 to December 2021. Before conducting the study, the participants read an information sheet and gave informed consent. If any further explanations were requested, the experimenter was at the disposal of the participant. The study was designed to the standards of the Ethical Review Panel of the University of Luxembourg and the Code of Ethics of the World Medical Association (Declaration of Helsinki). The study was conducted in line with the current COVID-19 guidelines.

Stimuli and Materials

The stimuli were Arabic numerals from 2 to 9. They were presented in Arial font in the center of a computer screen. The stimuli were presented in different colors. For the parity experiment, the numbers were presented in yellow and blue. For the magnitude experiment, the numbers were presented in green and red.

The experiment was programmed in PsychoPy3 v2020.2.8 (Peirce et al., 2019), running over Python (Python Software Foundation, USA). The computer used was an Acer Spin 3 laptop with a refresh rate of 60 Hz. The statistical analysis was run in SPSS Statistics 27 (IBM, USA). For the mathematical fluency test, five pages with math exercises were used.

Procedure

The main experiment is computerized. A series of numbers appear in different colors on the screen: yellow and blue in the parity experiment, or red and green in the magnitude experiment. The participant is asked to respond to the colors using the computer keyboard. Specific instructions are given before each experimental part.

For the parity experiment, the participant is asked to press on the down and up keys with their second and third fingers, respectively; for half the participants, up is for yellow and down is for blue. For the magnitude experiment, the participant has to press on the left and right keys, also with the second and third fingers. For half the participants, left is for green and right is for red. The participants are asked to use their dominant hand to respond in each experiment.

The participants were randomly divided in two equal groups: 17 participants in the experimental group and 17 participants in the control group. In the experimental group, numerical concepts of parity or magnitude are linked to the color, so for example even numbers like 4 would be associated with blue, while odd numbers would be associated with yellow. In the item-level control group, there is no link between the color and a numerical concept, so for example blue could stand for 2, 4, 3, 6 or 7.

An implicit learning paradigm is used to support the formation of color-number associations. In the experimental version of the parity experiment for example, in congruent trials, all the even numbers most often occur in blue, and all the odd numbers most often occur in yellow. Congruent trials are about 90% of trials; incongruent trials are about 10% of trials. So the participant might learn implicitly that the even numbers are often blue. A trial with a yellow 4, for example, is an incongruent trial, because 4 is even and therefore is usually blue.

There were 5 blocks of 110 trials in each the parity and magnitude experiments. Each trial consisted of a number presented in a color (yellow or blue; green or red) for 250 ms, followed by a blank screen until the response, with a maximum of 3.75s. After a response, there is a blank screen (750 ms-1.75s), followed by the next trial. Throughout each block, 5 to 15 congruent trials are shown, followed by 1 incongruent trial, for an average ratio of 10:1 congruent:incongruent trials.

In the 6^{th} block of each the parity and magnitude experiments, double-digit numbers were presented in black. In this explicit color report task, participants were asked to use their

non-dominant hand to respond, with different keys than in the implicit learning blocks. In total, the computerized experiment lasted about 40 minutes per participant.

After the computerized part of the experiment, a TTR (paper-and-pencil Tempo Test Rekenen) was administered. This is a 5-minute test for mathematical fluency. Participants were given 5 pages with arithmetic problems on each page. The pages contain the following operations: 1) addition; 2) subtraction; 3) multiplication; 4) division; 5) a mix of all these operations. The participant had one minute per page and was instructed to give as many correct solutions as possible. We planned to use this test to see if there would be a correlation between individuals' mathematics ability with their implicit association effects.

Results

The means and standard deviations of accuracy and response time (RT) for both the parity and magnitude experiments can be found in Table 1.

Table 1. Means and standard deviations (in parentheses) of accuracy (in proportion correct) and response time (in seconds).

		Parity		Magnitude	
		Experimental	Control	Experimental	Control
Accuracy	Congruent	0.890 (0.076)	0.907 (0.070)	0.896 (0.079)	0.917 (0.056)
	Incongruent	0.846 (0.088)	0.895 (0.082)	0.878 (0.102)	0.901 (0.090)
RT	Congruent	0.440 (0.082)	0.420 (0.071)	0.423 (0.053)	0.414 (0.074)
	Incongruent	0.438 (0.082)	0.425 (0.080)	0.439 (0.064)	0.417(0.083)

For the statistical analyses, we used repeated-measures analysis-of-variance (ANOVA) consisting of a within-participants factor of *congruency* (congruent vs. incongruent trials) and a between-participants factor of *group* (experimental vs. control). Follow-up *t*-tests to compare congruent and incongruent trials in each the experimental group and the control group were applied separately for accuracy and RT. These were one-tailed, paired-samples *t*-tests.

Parity Accuracy

For the Parity accuracy, we plotted a graph that indicates the standard error by showing the graph of means and the error bars. The graph shows an accuracy of 0.5 to 1. There was a 4.4% lower accuracy for incongruent trials, but little difference in the control group (Figure1).

There was a significant main effect of *congruency*, $F(1) = 12.33$, $p = .001$, $\eta_p^2 = .28$. The interaction of *congruency* and *group* bordered on significance, $F(1) = 4.05$, $p = .053$, $\eta_p^2 = 0.11$. The difference between incongruent and congruent trials was significant for accuracy in the parity experiment for the experimental group, $t(17) = 4.13$, $p < .001$, $d = 0.04$. This difference was not significant for the corresponding control group, $t(16) = 1.00$, $p = .17$, $d = 0.012$.

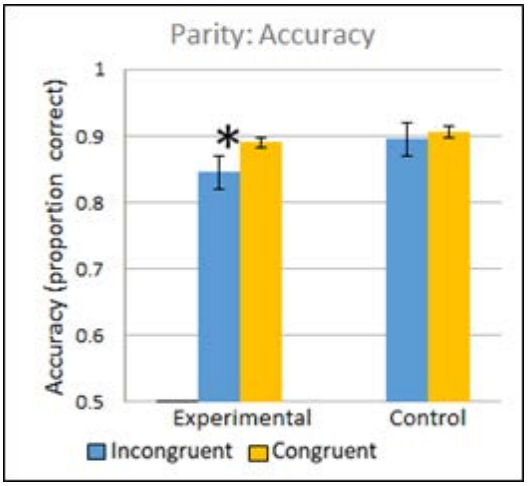


Figure 1. Accuracy for the parity (blue/yellow) experiment: bar graphs represent means and error bars represent ± 1 standard error.

Parity Response Time

For the Parity response time, we plotted a graph that indicates the standard error by showing the graph of means and the error bars. The graph shows an accuracy of 0.25 to 0.5. This graph indicates little difference in the control and the experimental group. In terms of Parity RT for the control group the incongruent parity RT is slightly higher than congruent.

There was no significant main effect of *congruency*, $F(1) = .24$, $p = .63$, $\eta_p^2 = .007$. The interaction of *congruency* and *group* was not significant, $F(1) = .24$, $p = .35$, $\eta_p^2 = .027$. The difference between incongruent and congruent trials was neither significant for accuracy for the experimental group, $t(17) = 0.35$, $p = .36$, $d = 0.0016$, nor for the corresponding control group, $t(15) = 0.93$, $p = .18$, $d = 0.005$.

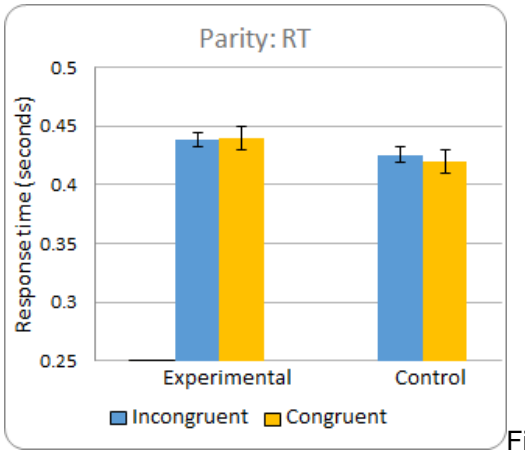


Figure 2. Response time (RT) for the parity (blue/yellow) experiment: bar graphs represent means and error bars represent ± 1 standard error.

Magnitude Accuracy

There was a significant effect of *congruency* $F(1)=4,52$, $p = 0.41$, $\eta_p^2 = 12$. The Interaction of *congruency* and *group* was not significant, $F(1) =0.24$, $p = .88$, $\eta_p^2 = .124$. However, the difference between incongruent and congruent trials was significant for accuracy in the experimental group, $t(17) = -1,90$, $p < .038$, $d = 1,83$. This difference was not significant for the corresponding control group, $t(15) = 1.205$, $p = .12$, $d = 0.016$.

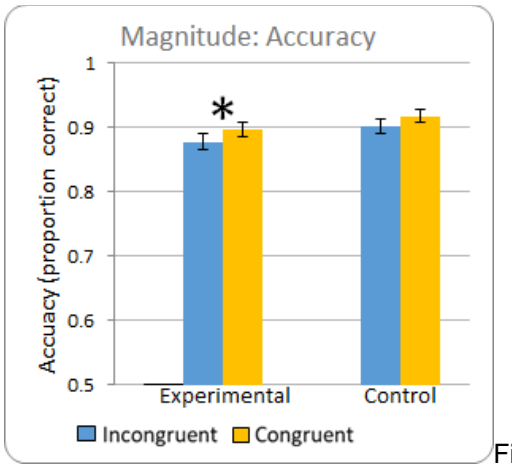


Figure 3. Accuracy for the magnitude (red/green) experiment: bar graphs represent means and error bars represent ± 1 standard error.

Magnitude RT

For the Magnitude response time, we plotted a graph that indicates the standard error by showing the graph of means and the error bars. The graph shows an accuracy of 0.25 to 0.5. In terms of experimental Group, the response time in incongruent trials was higher for a total of 15 milliseconds in comparison to the congruent trials. In the control group there were little differences.

There was a significant main effect of *congruency*, $F(1) = 5,28$, $p = 0.28$, $\eta_p^2 = .14$. The interaction between *congruency* and *group* was not significant, $F(1) = 2,63$, $p = .114$, $\eta_p^2 = 0.76$. The difference between incongruent and congruent trials was significant for the experimental group, $t(17) = 2.35$, $p < .015$, $d = 0.015$. This difference was not significant for the corresponding control group, $t(15) = 0.68$, $p = .25$, $d = 0.003$.

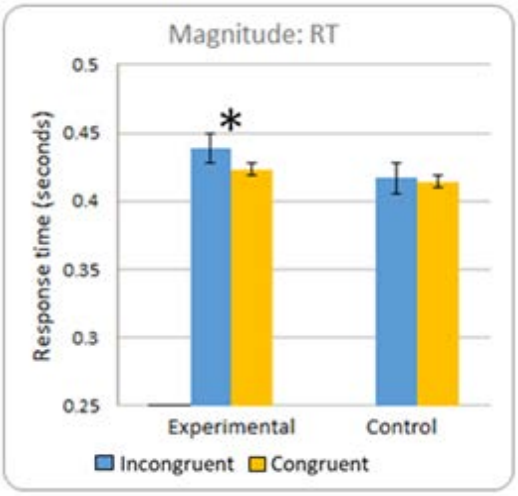


Figure 4. Response time (RT) for the parity (blue/yellow) experiment: bar graphs represent means and error bars represent ± 1 standard error.

Double-digit Accuracy

For the double-digit Accuracy, we plotted graphs that indicate an accuracy of 0 to 1 for parity and magnitude. The graphs for parity such as magnitude show no high accuracy. In terms of the experimental Group, there was a slight increase compared to the control group. The experimental group shows a 55% accuracy in terms of magnitude. The experimental group shows a 58% accuracy in terms of parity.

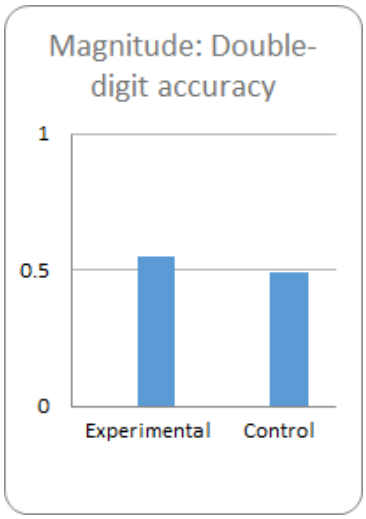


Figure 5. Double-digit accuracy for magnitude.

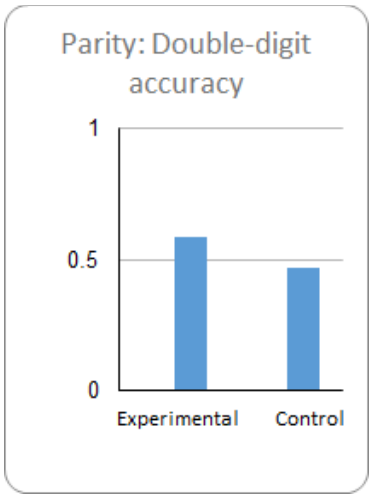


Figure 6. Double-digit accuracy for parity.

Discussion

Our experiment had the aim to analyze implicit associative learning of colors and numbers through high-probability associations. We used different computer-based tasks to see if it is possible to make an association between numerical concepts and the color they are most often shown in. In our study we looked at the concepts of parity and magnitude linked to the theoretical question of whether these concepts are processed automatically or not. We compared an experimental group, in which the concept and the color were consistent, to a control group, in which the colors were assigned to non-conceptual number groups.

Our results showed that the parity experiment had no significant differences between incongruent and congruent trials in response time. The accuracy, in the experimental group, for incongruent trials was lower than for the congruent trials, which is shown by the percentage of 4.4% difference, and which was a highly significant difference. These results can be considered as a small difference in accuracy but are highly reliable across participants. For the magnitude experiment, the response time of the experimental group was slower for incongruent trials than for congruent trials. This is a small but significant difference in response time (15 ms). Accuracy, for the experimental group, was lower for the incongruent trials than for the congruent trials with a percentage difference of

1.8%, which is a small, but significant difference. All our results were in the predicted direction, meaning that accuracy was lower and the response time longer for incongruent than for congruent trials. It should also be noted that for both the parity and the magnitude experiment there were no significant effects in the control group.

If we compare our results with other research papers that have analyzed similar topics we can see in general that the results of our study are a bit better than those of similar studies. For example, in the study "Category learning in the color-word contingency learning paradigm" from Schmidt et al. (2018) we can see that they had small effects in comparison to the ones we have for differential response time and accuracy. Their results for congruent trials were 5.1% and for incongruent trials 5.8% (which equals to only .7%-points difference). This can't be taken at face value because there are design differences between their and our studies: in our study, for example, the numbers were shown several times in the same colors, which was not the case in the study by Schmidt et al. because they showed the word stimuli only once, which again could be a reason for the smaller effect they have obtained. However, our study had fewer participants for the experimental group (N=18) but still highly significant results.

One thing to consider in interpreting these results is that our control group was not completely neutral, since there was also a higher probability for some numbers to be shown in the same color, since the same number-color pairings were shown more than once, but there was still no learning effect. This also shows that an association might not occur by chance, or even with a more subtle or less conceptual number-color grouping. Therefore, the effect in the experimental group is likely due to the conceptual experimental manipulation.

Another thing to consider is the impact of the task: our subjects had to do two experiments on the computer where they had to determine the color in which a presented number was shown (green/red or yellow/blue) by either clicking on the right/ left arrows or the up/ down arrows. After the computer-based experiments, each participant had to do a five page math test. The task in itself to say which color the number appeared in does not seem difficult, but the feedback of the subjects indicates that it was not as easy and that it could be a frustrating task. This is mostly not to be seen in our results, but there are some exceptions that can be explained, among others by a confusion of the keys during the experiment. In other words, this means that the overall accuracy is high. Each number was displayed for a duration of 250 ms. Some of our subjects stated that the computer task was the most difficult one to do because of the rapidity they had to determine which color the number was, which they explained with the fact that their body would react faster than their perception. It seemed to them as if their finger would press a response key before their brain even processed the color. A longer time to respond between the numbers would have been easier for the subjects and the experiment would have had an accuracy level that reached nearly 100% for the incongruent and congruent trials, however this might have also made performance for incongruent trials too easy to see an effect.

For future analysis or experiments it could be a good idea to separate the results of the subjects that noticed parity and magnitude from the subjects that did not and compare the results. Analyzing the correlation of the math test with the magnitude of the experiment effect could also be very interesting for future studies. This would be done separately for the accuracy or the response time effects. Another interesting analysis could be to examine whether there is an order effect. The participants in the current experiment learned an implicit association of color with either parity or magnitude first. Then they learned the other implicit association with different colors, but the same numbers. So a number that in the first part was associated with blue for parity, might later have been associated with green for magnitude. There might have been interference hindering the second association. Since in our experiment the number of participants who started with parity is the same as the number of participants starting with magnitude, we could do an analysis comparing the magnitude of the parity and magnitude effects depending on the order. For

example if the accuracy effect for parity was 4%, we could see if participants who did the parity experiment had a bigger effect than those who did parity second, which would mean that there might be a systematic difference around average, depending on the order. This research could have an important impact on further studies/applications, because if such an interference has an effect on the implicit learning, there could also be a more general interference in experiments associating number or words with colors, since numbers and words are usually displayed in black (on a screen) or blue (written by hand). Therefore the simple association with any color could also be subject to a later interference in implicit learning.

As far as our hypotheses are concerned, we have evidence in favor of our hypotheses. To be more precise, the null hypothesis – that people do not process parity and magnitude when their attention is on color and therefore there is no difference between the incongruent and congruent trials for the experimental group – can be rejected. Thus our results suggest that people do process numerical attributes such as parity and magnitude automatically, even when their attention is on color. We could still question the process of distraction of color information by parity because there is more evidence that this takes place by magnitude than for parity. Some literature even questions if parity is processed at all, if a task is not explicitly related to parity.. There were two possibilities as to which effect we thought might be stronger - parity or magnitude. It could be argued that magnitude might show a bigger effect because it is easier to notice and the recognition happens more naturally whether a number is big or small. On the other hand, the effect might have been bigger for parity, since it is an easy and well-known concept for participants. What also needs to be taken into consideration is that parity is categorical, so it makes sense to assign blue numbers to even numbers, for example - all even numbers (2, 4, 6, 8) are perfect examples of even-ness. Also, the numbers used in the main part of our experiment were one-digit numbers, and therefore it could be argued that all the numbers could be considered as small. Our results indeed showed that the effect in accuracy was bigger for parity than for magnitude. But since the numbers were all below ten, when comparing them to each other, magnitude was also an easy concept. This could explain why, even though the effect in accuracy was smaller for magnitude, our results showed an effect for magnitude in both accuracy and response time. Magnitude on the other hand is relative, so while there might be a big effect for 2 or 9, there might not be for 5 and 6. Some numbers are seen as less big or small than others, which shows that they aren't good examples of magnitude differences, which can have the consequence that color categories might not be learned as well.

One question raised in the scientific literature is the level at which such processing occurs, which is still debated. It could be at the sensory level (the color is associated with the shape of the numbers, early in the visual process, without any higher processing); the response level (the participants associate the colors to a response key and the motion of pressing that key with a certain finger, rather than with the number); or the conceptual level (across semantic categories, higher level association between the color and the numerical concept, rather than at an individual item level). In the current experiment four numbers were associated with one color, e.g. blue, while four other numbers were associated with another color, e.g. yellow. Thus, the sensory and response levels were well matched in both the experimental and control group and the types of responses were the same. If learning occurred in both groups, this might be evidence for a contribution of sensory and response level learning. Instead, there is no effect in the control group and learning occurred only for the participants of the experimental group, which can be interpreted as evidence that, in the experimental group, learning occurred at a conceptual level. This means that the participants implicitly learned about the parity and magnitude concepts which facilitated associative learning. Schmidt et al. (2018) also suggested that the association in their study between word categories and colors occurred at a conceptual level. In contrast to the current study, this experiment used word categories, for which the choice of categories and colors seems more

arbitrary. For example, there is no conceptual association of animal words with the color purple. When numbers are used, the chosen categories and control groups can be applied more systematically. Another important difference in the experimental design is that Schmidt et al. (2018) did not have a control group, in which the association of words and categories was completely arbitrary, and could have been compared to the experimental group.

What would be interesting is to have a look at where and how implicit conceptual expressions could be used, outside of the experimental setting, and also beyond numbers, since color may be used as help to discriminate between words, letters, or numbers, but can also serve as motivation, especially in early reading (Otto & Askov, 1968). In the TV-series Numberjacks there was a color given to each number randomly and some of the colors resemble each other, but not systematically. In their study, Rinaldi et al.(2019) showed that the number-color association of Numberjacks had no measurable effect on the children's learning: children who watched the show had no learning advantage in numerosity or mathematical tasks over those who did know the color association from it. A probable explanation is that the associations were made at random and thus had no helpful correspondence to conceptual categories.

It might be more helpful to assign colors to numbers in a more meaningful way. In her experiment on the Cuisenaire rods experiment, Hater (1970) assessed the Cuisenaire rods, in which numbers were sorted into categories (number families) and a color family was applied to each number family in a meaningful way, which is how Georges Cuisenaire developed the Cuisenaire rods in 1952. They were meant as an approach to introduce mathematics at the secondary school level (De Bock et al., 2020). The numbers three, six and nine were associated with different shades of green and blue; the numbers two, four and eight in purple and brown shades; and the numbers five and ten in warm colors of red and orange. The relationship between the magnitude of the rods and the colors followed a categorical concept, which is more intuitive than a random association and might therefore have a stronger effect on learning.

Using concepts might also make learning easier as it reduces the amount of items to learn. To be more precise about why the Cuisenaire rods are useful, when learning the numbers of the Cuisenaire rods, there are only three number families and three color families to learn, whereas if the concept of families is not used, nine individual numbers and nine individual colors would have to be learned. Also, some numbers are more difficult to learn as belonging in the same category than others, since numbers are not completely neutral stimuli. For example, two and nine are different in parity and magnitude, which might create an "intuitive" resistance to put them into the same category.

An example for meaningful associations is that due to "Hebbian" co-activation, neural responses to the color blue might be correlated with neural responses to even numbers through neural associations. This is a possible explanation as to why the association of the color blue with even numbers is easier to learn and could be used as a meaningful association when learning. It could be argued that there is also Hebbian learning for individual items, but it might work better at a conceptual level, because the color blue might not be associated with single even numbers, but with the concept of parity. Even though the effects we measured were at a behavioral level, it might be interesting to observe what happens at a neural level in future studies. The participants of our study learned the number-color association implicitly, but the connection might also have taken place as an unconscious neural response, which might then feel like an intuitive association of numbers with certain colors. An EEG study could help determine whether, when learning of implicit color associations occurs, it is seen first as a neuronal or behavioral response.

Some studies supported the theory in the learning field, that using color associations in learning material is not useful or might even have a negative impact on learning.

Skulmowski (2021) argues that even though using color codes may reduce the cognitive load and thereby simplify complicated visualizations and facilitate learning, color codes can also hinder learning if the learner becomes dependent on them for recall. In such a case, color codes would only still be useful if they are also used in the test or recall tasks later on. The study by Rinald et al. (2020) had the results that color associations for Numberjack colors had no significant impact on the children's learning and was therefore not useful. This non-useful or even negative impact of color association occurs when the color is not applied in a systematic or conceptual way, like for the participants of our experiment in the control group (for those participants no implicit learning occurred).

Color associations can be used strategically (for numbers, animals, word categories, etc.), like with the use of text markers. People generally use them according to their own planned association, for example marking all new vocabulary in one color. This influences their perception of those words and influences their memory. For example, when learning a new language, new vocabulary could be color-coded systematically and consistently to facilitate later recall. Thus, the implicitly learned color association would have a positive effect on memory. Some approaches make use of this positive effect. On the other hand, if that color-code is "broken," e.g. by using a textbook with different colors, the previously learned association could interfere with the new learning. The question in which context color associations could be useful and in which it could be harmful could be an interesting topic for further research. Also the duration of such implicit color associations plays a role. It seems unlikely that color associations would last long-term. If the same participants re-took our experiment a few weeks later, we would not expect them to still have the same implicit color association that they learned the first time. But when teaching children parity, a long-term implicit color-association could prove useful and facilitate learning. Once the children acquire the knowledge of parity, the color association might become useless and be forgotten.

Literature

ATHANASOPOULOS, P. (2009). Cognitive representation of colour in bilinguals: The case of Greek blues. *Bilingualism: Language and Cognition*, *12*(1), 83–95. https://doi.org/10.1017/s1366728908003 88x

Bankieris, K. R. & Aslin, R. N. (2016). Implicit associative learning in synesthetes and nonsynesthetes. *Psychonomic Bulletin & Review*, *24*(3), 935–943. https://doi.org/10.3758/s13423-016-1162-y

De Bock, D. (2020). Georges Cuisenaire's numbers in colour. A teaching aid that survived the 1950s. In *"Dig where you stand" 6. Proceedings of the sixth International Conference on the History of Mathematics Education*. WTM-Verlag; Münster.

Ellis, C. (2006). Numberjacks. British Broadcasting Company. Retrieved from http://www.bbc.co.uk/programmes/b006mhcr

Hater, M. A. (1970). Investigation of color in the Cuisenaire rods. *Perceptual and Motor Skills*, *31*(2), 441-442.

Lin, O. Y. H. & MacLeod, C. M. (2018). The acquisition of simple associations as observed in color–word contingency learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(1), 99–106. https://doi.org/10.1037/xlm0000436

Numicon (Oxford University Press, 2018). (2018). *Numicon*.

Numicon (Oxford University Press, 2018). (2018). *Numicon.*

Otto, W., & Askov, E. (1968). The role of color in learning and instruction. *The Journal of Special Education*, *2*(2), 155-165.

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Richard Hochenberger, Sogo, H., … Jonas Kristoffer Lindelov. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Purves, D. (2019). *Brains as engines of association: an operating principle for nervous systems*. Oxford University Press.

Reynvoet, B., Caessens, B. & Brysbaert, M. (2002). Automatic stimulus-response associations may be semantically mediated. *Psychonomic Bulletin & Review*, *9*(1), 107–112. https://doi.org/10.3758/bf03196263

Rinaldi, L. J., Smees, R., Alvarez, J., & Simner, J. (2019). Do the Colors of Educational Number Tools Improve Children's Mathematics and Numerosity? *Child Development*, *91*(4).

Schmidt, J. R., Augustinova, M. & de Houwer, J. (2018). Category learning in the color-word contingency learning paradigm. *Psychonomic Bulletin & Review*, *25*(2), 658–666. https://doi.org/10.3758/s13423-018-1430-0

Schmidt, J. R., Crump, M. J., Cheesman, J., & Besner, D. (2007). Contingency learning without awareness: Evidence for implicit control. *Consciousness and Cognition*, *16*(2), 421–435. https://doi.org/10.1016/j.concog.2006.06.010

Skulmowski, A. (2021). When color coding backfires: A guidance reversal effect when learning with realistic visualizations. *Education and Information Technologies*, 1-16.

Statista. (2018, 7. Juni). *Häufigkeit der Verwendung von Textmarkern*. Abgerufen am 12. Januar 2022, von https://de.statista.com/statistik/daten/studie/178514/umfrage/haeufigkeit-der-verwendung-von-textmarkern/

Interoception and the menstrual cycle

Diana BÄRENZ, Vanessa JOACHIM, Chiara JUNK, Silas-Joy KIEFER, Mara MICHELS, Alana STURGEON

Supervised by Dr. Angelika Dierolf and Dr. Marian van der Meulen

The purpose of this article is to examine the interoception and perception of experimentally induced pain across the menstrual cycle of a group of 14 healthy females. Each woman was tested at three points in their menstrual cycle, the ovulation, the follicular, and the luteal phase. We assessed interoceptive awareness, accuracy and sensitivity, as well as pain sensitivity in form of pain threshold and pain tolerance measurements with a variety of different tests, such as the Schandry-task, the Thermode-Test and the Cold-Pressor-Test. Different studies found effects on pain perception between the different cycle phases. The results suggest that the menstrual cycle has an effect on the measured interoceptive concepts and pain perceptions and therefore their pain sensitivity. The cycle had a significant effect on interoceptive accuracy and the heat pain thresholds whereas it had no significant effect on the remaining interoceptive concepts, the cold pain threshold and the pain tolerance. They further imply that overall fitness and mental health have a relationship and an effect on the interoceptive concepts throughout the menstrual cycle phases. In conclusion, interoceptive accuracy changes throughout the menstrual cycle, but since our sample size is rather small and the existing literature is very limited as well as contradicting, we suggest further research on this topic.

Introduction

Interoception can be defined as the perceptions from inside the body, this includes the perception of physical sensations related to internal organ function such as heartbeat, respiration, satiety, as well as the autonomic nervous system activity related to emotions (Vaitl, 1996; Cameron, 2001; Craig, 2002; Barett et al., 2004). Furthermore, interoception includes two different forms of perception: proprioception and viscероception. Pain perception is a form of proprioception, where signals from the body are received primarily from the skin and the musculoskeletal apparatus (joints, tendons, muscles). Viscероception (Latin: viscera=inner organs) is the term used to describe signals arising from the inner organs (Vaitl,1996). Interoceptive signals are transmitted to the brain via multiple pathways and there are multiple networks underlying interoception (Migliorini, 2017). The primary interoceptive representation in the cerebral cortex of the brain (dorsal posterior insula) engenders distinct, highly resolved feelings from the body which include pain, temperature, itch, sensual touch, muscular and visceral sensations and many more (Craig, 2002). Additionally, interoception can be divided into three dimensions, (1) interoceptive accuracy (reflected for example by the performance on objective behavioural tests of heartbeat detection), (2) interoceptive sensibility (self-evaluated assessment of subjective interoception using interviews/questionnaires) and (3) interoceptive awareness (metacognitive awareness of interoceptive accuracy (e.g., confidence-accuracy correspondence regarding participants answers on a heartbeat perception task). Since this sensibility (2) can be assessed by using subjective measures from the individuals' belief in their own interoceptive ability and the degree to which they feel engaged by interoceptive signals in the form of a self-report, it might reflect biases in subjective thresholds irrespective of interoceptive accuracy. To counteract this, it is recommended to use combined measures such as a heartbeat perception task performance with a measure of subjective confidence in performing the task (Garfinkel, 2015; Ehlers et al., 1995). Overall, interoception covers a broad range of sensations and for instance because of its close

relationship with psychiatric illnesses it is of great interest for numerous research projects. In our study we focus on the three dimensions of interoception, proprioception in form of pain perception from the skin and viscероception in form of the heart, since heartbeats are distinct and frequent internal events that can easily be discriminated and measured (Garfinkel, 2015). Another bodily sensation women experience regularly is the menstrual cycle. Many women report a difference in the perception of their body throughout the menstrual cycle (Silberstein & Merriam, 2000). However, the previous research on interoception and the menstrual cycle is extremely limited to almost non-existent. Consequently, we want to have a closer look at the menstrual cycle in connection with interoception throughout our study.

Firstly, a common way to measure interoceptive accuracy (1), is the Schandry task, which requires an individual to count their perceived heartbeats during a specific time period ("Heartbeat Tracking" e.g., Schandry, 1981). For the interoceptive sensibility (2) we use a self-report questionnaire, the MAIA-II (Mehling et al., 2012). As stated earlier, pain perception can be seen as proprioceptive interoception and the existing research gives an insight on possible changes in the menstrual cycle. Different studies found effects on pain perception between the different cycle phases (Procacci et al.,1974; Goolskasian, 1980, 1983; Riley et al., 1999; Fillingim & Maixner, 1995). Research has shown that pain sensitivity changes during the menstrual cycle in humans and animals, which has sometimes been ascribed to hormonal variations (de Tomaso, 2011; Hellström & Lundberg, 2000). Some studies report greater pain sensitivity during the premenstrual phase at ovulation or during menses (Goolskasian, 1980, 1983; Hellström & Lundberg, 2000). Other studies found a higher threshold during the follicular phase compared with the other phases (Riley et al., 1999). This just shows us how contradicting the literature itself is on the topic regarding the impact of the menstrual cycle on pain.

In addition, a study has found that trained dancers have increased interoceptive accuracy (Christensen et al., 2018). Dancers that have been training for years, had a higher interoceptive accuracy than junior dancers or the control group, this was measured by using a self-report questionnaire and a heartbeat perception task with a following confidence rating by the participants (Christensen et al., 2018). To conclude, sportif activity, especially dance seem to have an impact on interoception. There was another study that analysed if integrative exercise using aerobic and resistance exercise in mindful-based principles and yoga would help war veterans with the treatment of war-related post-traumatic stress disorder (PTSD) (Mehling et al., 2017). The interoceptive awareness was measured using a self-report questionnaire, the MAIA (Mehling et al., 2012). The results reported significant improvements in mindfulness and interoceptive bodily awareness in war veterans with PTSD (Mehling et al., 2017). In summary, exercises based in mindfulness such as yoga could have an impact on interoception. In the context of these findings, it would be interesting to further research if there are any correlations between interoception and personal fitness. The IFIS (International Fitness Scale) gives insight into one's physical fitness abilities and awareness in the form of a self-report questionnaire. However, there is no research in connection with fitness and interoception and the menstrual cycle, but exercise is commonly cited as a remedy for menstrual symptoms (Sutar et al., 2016).

Feedback from the body is assumed to be altered in depression and one of the many symptoms of depression can be numbness. Another study analysed interrelations between the ability to perceive heartbeats accurately, depressive symptoms and anxiety in healthy participants (Pollatos, 2009). As a main result they observed a negative correlation between heartbeat perception and depression. But only when focusing on high anxiety levels this negative correlation remained significant. In conclusion, there is a possible relationship between depressive symptoms and interoceptive awareness, but further research is necessary (Pollatos, 2009). In addition, another study researched the pre-post effect of cognitive-behavioural therapy (CBT) in a depressive

sample, since studies have shown that CBT and mindfulness interventions are promising approaches to improve interoceptive abilities and interoceptive accuracy and sensibility are decreased in depressive samples (Karanassios et al., 2021). As a result, the depressive sample showed a significant decrease in depressive symptoms and increased mindfulness and interoceptive abilities after CBT. Again, the research in connection with the menstrual cycle is extremely limited. For example, on a retrospective menstrual distress questionnaire women said that they experience increases in anxiety, irritability, depression and tension in the premenstrual phase of the cycle (Parlee, 1982). Additionally, hormonal contraception may be associated with depressive women, especially among young adolescents but evidence did not support that hormonal contraception directly causes depressive symptoms (Buggio, 2021). A different study found that oral contraceptives do not alter the physiological fluctuation of mood occurring with menses in young healthy women (Natale & Albertazzi, 2006).

Hypotheses

In this study, our aims are to investigate 1) How does interoception change throughout the menstrual cycle and 2) How does pain perception change throughout the menstrual cycle?
Just like some previous studies, we use elements such as the Schandry task or the MAIA-II to expand the findings on interoception and the menstrual cycle with already existing research, as well as our own questions and aspects. Since there are no consistent results on the direction of changes between the phases in the literature, we have no strong hypotheses about the direction of any potential differences across the menstrual cycle in interoception or pain perception in our study.
Therefore, new hypotheses as well as very general formulated hypotheses were formed.

1. We expect a difference in interoception between different menstrual phases.
2. We expect different levels of sensitivity in pain perception during the different menstrual phases.
3. Fitness and depression correlate with interoceptive sensibility, awareness and accuracy and differ in the different menstrual cycle phases.

Methodology

Participants

For our study we recruited participants via social media, an online student platform of the University of Luxembourg and word of mouth. We could not include everyone in our study, who was interested in participating and contacted us, because they also had to fit with our recruitment criteria. Therefore, our exclusion criteria were as follows: being pregnant, being in the menopause, having an Implanon and getting the injectable depot contraceptive. Participants who took the birth control pill or the progestin pill without a monthly break were excluded too, unless they got their monthly period, nevertheless. Having an irregular cycle was an additional exclusion criterion, meaning their period did not have a recurring interval of 22 to 35 days, plus or minus three days variations, while the duration of menstruation should not exceed 14 days or fall below three days. We also decided to exclude participants with a recent change of the contraceptive method they use. This measure was taken in order to avoid any bias regarding a messed up hormonal balance as a consequence of recently changing hormonal contraceptives. Based on our exclusion criteria and the participants' responses in the pre-screening, we had to exclude nine potential participants. Five of them could not be included because of their very irregular menstrual cycle and not having their menstrual bleeding every month. Another potential participant did not fit with our recruitment criteria, as her menstrual cycle usually lasts 40 days, and while the duration of another interested participant's menstruation falls below three days, she could not participate either. We also had to exclude two other young women, who contacted us, because they had recently changed their contraceptive method from the hormonal birth control pill to using condoms, and they were still in the phase of adjustment, where the natural

hormonal balance must be recreated again. Nine women were excluded because their calculated lab sessions did not fit in our testing schedule. The consumption of alcohol or taking pain relievers 24 hours prior to the lab sessions was an exclusion criterion too, in four cases this was met. However, we still included those participants in our analyses. Therefore, our sample size is restricted to 14 participants in total (N = 14). 100% of the participants are female. We remunerated them with 25€ and two and a half participant hours.

The age of all participants varies between 18 and 27 years, while 92,9% are between 18 and 23 years old (N = 13) and one participant is between 23 and 27 years old (N = 1). One half of the sample indicated that they are in a (committed) relationship and the other half indicated that they are single at the moment. 92,9% were psychology students and all of them were fluent in German.

All participants had a regular cycle according to the criteria above, while the duration of their menstrual bleeding is shorter than one week for 12 of our participants (85,7%), and only two of them (14,3%) have their menstruation for up to 14 days. The range of their cycle lengths in general varies between 22 and 35 days, the mean for their minimal lengths of their cycles is 26.07 and for their maximal lengths is 29.57.

Looking at the contraceptive methods, three of the participants indicated that they do not use any contraceptives at all. Of those who use contraception, one participant takes the contraceptive pill for five years and another one for one and a half years now, but with a monthly break, so they still have their monthly period. Six participants use condoms as their contraceptive method; one of them in the past six months, three participants for six months up to a year, while two of them changed from the intake of the contraceptive pill to using condoms, another participant uses condoms for three years now. There is also one participant, who indicated that she has been taking the contraceptive pill for four years now, since she started using contraceptives, and she also uses condoms often as additional contraceptives. Two participants stated that they got a copper spiral or copper chain: one of them for between six months up to a year and the other one for one and a half years. Another participant indicated that she got a copper ball for six months up to a year now.

Two participants have regular medication intake. One participant takes Euthyrox – 0,075 mg (a thyroid hormone) and one participant takes Venlafaxine – 150 mg (remedy against depression and anxiety).

Two of our participants have a severe score (21-27) on the depression scale of the DASS, on the stress scale it is only one person with a severe score (26-33). For the anxiety scale, two participants have an extremely severe level (>20) (s. table 1) There was one person reporting a high score on the anxiety and the depression scale.

Table 1: Frequencies of Severity Levels in the Depression-Anxiety-Stress-Scale (DASS) (*N*=14)

DASS- Depression		DASS- Anxiety		DASS- Stress	
Severity	*N*	Severity	*N*	Severity	*N*
Normal (Score 0-9)	10	Normal (Score 0-7)	9	Normal (Score 0-14)	9
Mild (Score 10-13)	1	Mild (Score 8-9)	1	Mild (Score 15-18)	2
Moderate (Score 14-20)	1	Moderate (Score 10-24)	2	Moderate (Score 19-25)	2
Severe (Score 21-27)	2	Severe (Score 15-19)	0	Severe (Score 26-33)	1
Extremely severe (Score 28+)	0	Extremely severe (Score 20+)	2	Extremely severe (Score 34+)	0

Shows the frequency of every severity-level in the subcategories of the Depression-Anxiety-Stress-Scale (DASS)

Table 2.5: Descriptive statistics of the Depression-Anxiety-Stress-Scale (DASS) from the Pre-Screening (*N*=14)

Main variables	Measured concept	*M*	*SD*
Depression-Anxiety-Stress-Scale (DASS)	3 Subcategories: 1. Depression, 2. Anxiety, 3. Stress		
	Depression	7,14	7,75
	Anxiety	7,71	8,03
	Stress	11,71	8,85

Shows the mean and standard deviation of the Depression-Anxiety-Stress-Scale (DASS) from the Pre-Screening

Materials & Session procedure

Our study consisted of two parts, an online prescreening and a laboratory part performed at three fixed dates during the cycle. The prescreening was sent to them by e-mail beforehand along with the informed consent and ethics sheet, so they could fill it out at home. It amounted to approximately 20 minutes and contained the following questionnaires: FEM (Berliner Fragebogen zum Menstruationserleben) (Saupe ,1987), IFIS (International Fitness Scale) (Pereira et al., 2020), DASS-21 (Depressions-Angst-Stress-Skalen - deutschsprachige Kurzfassung) (Nilges & Essau, 2021), PSQ (Perceived Stress Questionnaire) (Fliege et al., 2001). In addition to that we asked them demographic questions and a thorough questioning of the cycle, symptoms related to the cycle and health questions. In our analyses we will focus on the results of the DASS and the IFIS.

If none of the exclusion criteria were met, we used the information about the cycle length and the date of onset of the last menstruation to calculate the dates when the participants should come to the laboratory. Therefore, we used the ovulation calculator from mummylator (*Eisprungrechner—Jetzt exakter Eisprung berechnen!* (o. J.)).

During the luteal phase we had to delay ten appointments in a range from three days prior to four days after the calculated date. In the follicular phase there were seven adjustments from two days before the calculated date to five days after it. In the ovulation phase we had to move six appointments in a range of five days prior to ten days after the calculated date. We had to delay six lab sessions because their menstruation did not start on time, in eight cases the laboratory was occupied. Seven appointments were moved due to the calculated date being on a weekend and three appointments due to scheduling problems. (s. table 2.6)

The date for testing during the luteal phase was five days before the start of menstruation and the date for testing during the follicular phase was on the third day after the start of menstruation. The date for testing during the ovulation phase was the calculated day of ovulation itself.

Each of the three laboratory appointments lasted 45 minutes and included the same procedure. At the beginning, the covid sheet was signed and online questionnaires were filled out again, in addition to several state questions, these were the MAIA-II (Mehling, 2018) and the FAW (Fragebogen zum aktuellen körperlichen Wohlbefinden) (Frank, 2011).

The 37 items of the MAIA-II are split into eight subscales, each statement deals with interoceptive awareness. You determine the arithmetic mean for each subscale to analyse this questionnaire.

Subsequently, we performed three interoception tests in the test cabin: the Schandry-task, the Cold Pressor test (Stening et al., 2007) and a thermode pain threshold. During these tasks, the participant was continuously connected to an ECG, so we could measure their heart rate.

We started with the Schandry-task, it consisted of three parts. Firstly, the participants had two resting periods, one with their eyes closed and one with their eyes open, each period lasting 160 seconds, the order was counterbalanced across participants. Secondly, the participants counted how many heartbeats they felt during four determined counting phases, previously there was one trial run. The trial run lasted 25 seconds, followed by four actual counting phases lasting 25, 35, 45 and 55 seconds, presented in a randomized order. So, in total the duration of the counting phases was 215 seconds, in which the participants had to pay attention to their interoception of their heart rate. After every counting phase they had to type in, how many heartbeats they had counted, followed by a confidence rating about their answer. In the third part, the participants counted the seconds during determined counting phases. The phases also lasted 215 seconds, with a trial run of 25 seconds and the counting periods lasting 25, 35, 45 and 55 seconds, again presented in a randomized order. The

participants typed their answer and gave a confidence rating.

Afterwards, the second and the third test were a Cold Pressor Test (CPT) and a thermal cold and heat pain threshold measurement with a thermode. The order of these two tests was counterbalanced across the participants.

The CPT was used to measure the pain perception and tolerance threshold of the participant. In this task we asked them to put their right hand in approximately 5 degrees Celsius cold water for no more than three minutes and evaluate the subject pain intensity every 20 seconds on a numeric rating scale from zero to ten (0= no pain; 1 = pain threshold; 10 = the highest amount of pain they could imagine in this situation). The participants were able to take her hand out of the water at any point, when the pain was no longer bearable. For taking a closer look at the heart rate of the participant during the CPT, one experimenter set a marker in the ECG recording file as soon as the participant put her hand in and when she pulled it out again.

For the other pain perception test we fixed the thermode to their left forearm and set three heat and three cold stimuli. The baseline temperature of the thermode was 32 degrees, it went up to a maximum of 50 degrees Celsius and down to a minimum of 0 degrees Celsius. The temperature change rate was 1.5 degrees per second for both the heat and the cold trials. The participants could have ended the stimuli as soon as they perceived any kind of pain, in order to determine their threshold. The return rate after their response was 8 degrees per second for heat and 4 degrees per second for the cold trials. Contrary to the CPT the participants should not have tried to stand the pain as long as they could. Instead, we wanted to know the time point when it just started to become painful for them. We asked them to press the “N”-Button on a keyboard right in front of them as soon as their threshold was reached and by pressing the button the temperature of the thermode was immediately going back to its starting point (32 degrees). During this test we wanted to have a closer look at the heart rate again and for this reason, one experimenter set a marker in the ECG recording file, when the test started with the first hot stimulus, between the third hot stimulus and the first cold one, and when it stopped after the third cold stimulus The participants could have ended the stimuli as soon as they perceived any kind of pain, in order to determine their threshold. Contrary to the CPT the participants should not have tried to stand the pain as long as they could. Instead, we wanted to know the exact temperature as soon as it started to become painful for them.

Statistical Analysis

All data analyses were performed with SPSS. The significance level α was set at 0.05. We looked at descriptive analyses such as mean values, standard deviations, and frequency analyses.

The interoceptive accuracy score was calculated by comparing the score of actual heartbeats to the score of counted heartbeats for each participant and later determining the mean value of the differences. By calculating the mean for each subscale of the MAIA-II-Questionnaire, the score for the interoceptive sensibility was determined. The mean score of the Schandry confidence task was calculated as the scores for the interoceptive awareness. To examine the relation between interoceptive accuracy, sensibility, and awareness and the menstrual cycle phases we conducted repeated measurement ANOVAs.

The means for the heat and cold thresholds of the thermode threshold test were calculated by conducting the means of the heat and cold thresholds of every session for each participant, which allowed us to determine the means for each menstrual cycle phase. For the Cold-Pressor-Test we calculated the means of the durations for the values of pain tolerance. Again, we conducted repeated measurement ANOVAs to examine the interaction of the menstrual cycle phases and pain sensitivity, once with the pain tolerance of the Cold-Pressor-Test and with the heat and cold pain thresholds of the thermode threshold test.

To examine the relation between fitness, measured by the International Fitness Scale, short IFIS, and interoception we conducted multiple bivariate correlation analyses. For interoceptive accuracy we correlated the IFIS scores with the Schandry scores, for interoceptive sensibility with the MAIA-II-scores and for interoceptive awareness with the scores of the Schandry confidence task.

The score of the DASS depression subscale was calculated by adding up the sum value of the corresponding items. With this score and the corresponding values for the subcategories of the interoception as described above, we conducted multiple bivariate correlation analysis, to examine the relation of interoception and depression.

Results

An overview of our main variables, the concept they have measured, and their respective means and standard deviations can be found in table 2.1 to 2.5 in the Appendix.

For our first hypothesis, repeated measures ANOVAs were conducted that examined an effect of the menstrual cycle on interoceptive accuracy, sensibility, and awareness. Since our sample is very small, we decided not to exclude the participants which indicated to have taken pain medication. However, we present both results. The results excluding the participants taking pain medication will be in presented in square brackets ([ ]) A significant interaction was found between the menstrual cycle phases and interoceptive accuracy $F(2,26)= 3.49$, $p = .046$, $\eta_p^2 = .211$ [$F(2,18)=4.53$, $p= .026$, $\eta_p^2= .335$]. Furthermore we found a significant linear increase from the luteal phase towards the ovulation $F(1,13)= 5.32$, $p= .038$, $\eta_p^2= .290$ [$F(1,9)=8.39$, $p= .018$, $\eta_p^2=.482$] (see Figure 1). The lowest scores were found in the luteal phase $M = .412$, $SD= .066$ [$M = .418$, $SD = .084$] indicating a low interoceptive accuracy.

It was found however that the menstrual cycle phases had, according to the self-report, no significant effect on the interoceptive awareness $F(2,26)= .259$, $p = .774$ [$F(2,18)=.511$, $p= .608$],measured by the confidence ratings in the Schandry task, or the two way interaction of the menstrual cycle and the MAIA-II-Questionnaire scales $F(14,182) = .663$, $p = .808$ [$F(14,126)= .540$, $p= .905$] which measure interoceptive sensibility.

For the second hypothesis also, repeated measures ANOVAs were conducted which examined the interaction between the menstrual cycle phases and pain sensitivity regarding heat and cold pain thresholds as well as pain tolerance. The results of the heat pain threshold revealed a marginally significant effect on the perception in terms of heat induced pain with a higher sensitivity during the ovulation

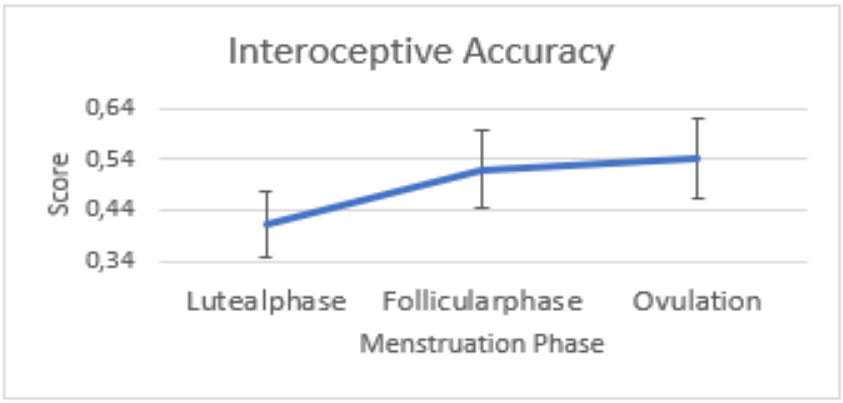


Figure 1: Interoceptive Accuracy measured by the Schandry task throughout the menstrual cycle phases. (*N*=14)

Error Bars: For the calculation of the error bars the standard error (SE) was used.

phase $F(2,26)= 3.15$, $p = .06$, $\eta_p^2= .195$ [$F(2,18)= 2.59$,$p= .102$, $\eta_p^2= .224$]. We observed a significant linear decrease in the pain threshold over the three phases $F(1,13)= 8.74$, $p = .011$, $\eta_p^2= .402$ [$F(1,9)=8.844$, $p=.016$, $\eta_p^2=.496$] (see Figure 2).The highest pain thresholds were found in the Luteal phase $M= 46.59$, $SD= .58$ [$M = 46.56$, $SD = .725$] indicating a low pain sensitivity in the cycle phase, Then decreasing through the menstruation phase, having a medium threshold $M=46.13$, $SD= .065$ [$M = 45.93$, $SD =.832$]and the ovulation $M= 45.74$, $SD = .647$ [$M = 45.62$, $SD = .68$] showing the lowest threshold. (s. Figure 2)

No significant interaction between the menstrual cycle phases and cold pain sensitivity (thresholds) could be found $F(2,26)= .702= p= .505$ [$F(2,18)=2.49$, $p= .111$, $\eta_p^2= .217$. We found a marginally significant linear increase from the luteal phase towards ovulation $F(1,9)= 3.74$, $p= .085$, $\eta_p^2= .293$].

With the Cold Pressor Test, the construct of pain tolerance was measured by the time the participants submerged their right hand in ice-cold water. To reduce the risk of injury, we limited the time to 180 seconds. The menstrual cycle phases had no significant effect on pain tolerance $F(2,26)= .175$, $p= .840$ [$F(2,18)=1.357$, $p= .282$].

Since nine of the 14 participants kept their hand submerged in the cold water for the limit of 180 seconds, we were able to observe a ceiling effect (s. Figure 3).

For our third hypothesis we examined whether there is a correlation between the concepts of interoception, interoceptive accuracy, interoceptive awareness and interoceptive sensibility under consideration of the menstrual cycle, and the fitness, measured by the International-Fitness-Scale, short IFIS, as well as depression measured by the Depression-Anxiety-Stress-Scale (DASS). To analyse this, we did multiple bivariate correlational analyses.

Starting with the correlations between interoceptive accuracy and fitness, we found significant positive correlations between the Schandry scores and the IFIS subcategory “muscular strength” for the follicular phase $r(14)= .463$, $p= .048$ (s. Figure 6.1) and the ovulation $r(14)= .469$, $p= .045$ (s. Figure 6.2) (s. Table 3.1).

Moving on to the correlations between interoceptive awareness, measured by the

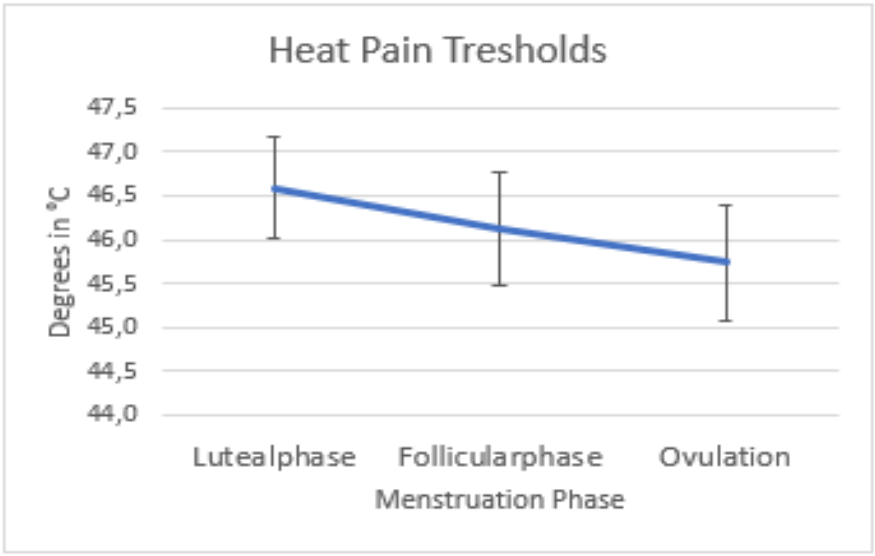


Figure 2: Pain perception: Mean heat pain thresholds measured with the Thermode test throughout the menstrual cycle phases. (N=14)

Error Bars: For the calculation of the error bars the standard error (SE) was used.

Schandry-confidence-task, and fitness, we found no significant correlation. (s. Table 3.2)

Table 3.1: Correlations Schandry-scores with International Fitness Scale (IFIS) (*N*=14)

		IFIS_General Physical Fitness	IFIS_cardiorespiratory Fitness	IFIS_muscular strength	IFIS_Speed/Agility	IFIS_Flexibility
Luteal phase	Schandry	.085	.037	.390	.127	-.050
Follicular phase	Schandry	-.094	.145	.463*	.124	-.170
Ovulation	Schandry	.220	.245	.469*	.334	.055

*Correlation is significant at the 0.05 level (1-tailed)

Table 3.2: Correlations Schandry-Confidence with International Fitness Scale (IFIS) (*N*=14)

Menstruationphase	Variable	IFIS_General Physical Fitness	IFIS_cardio-respiratory Fitness	IFIS_ muscular strength	IFIS_Speed/ Agility	IFIS_ Flexibility
Luteal phase	Schandry Confidence	-.016	-.052	-.019	.006	-.021
Follicular phase	Schandry Confidence	.011	.186	-.013	.236	.013
Ovulation	Schandry Confidence	.140	.221	-.013	.224	-.006

Table 3.3: Correlations MAIA-II with International Fitness Scale (IFIS) (*N*=14)

Menstruation phase	Variable	IFIS_GeneralPhysicalFitness	IFIS_cardiorespiratoryFitness	IFIS_muscular strength	IFIS_Speed/ Agility	IFIS_Flexibility
Luteal phase	MAIA_Bemerken	.189	.639**	-.048	.631**	-.037
	MAIA_NichtAblenken	.066	.347	-.268	.315	-.224
	MAIA_KeineSorgenMachen	-.055	.399	-.154	.307	-.050
	MAIA_AufmerksamkeitsSteuerung	-.281	.368	-.299	.291	-.117
	MAIA_EmotionalesGewahrsein	-.137	.250	.072	.541*	.237
	MAIA_Selbstregulation	-.131	-.237	-.186	-.378	-.419
	MAIA_AufLeibHören	-.153	.215	-.060	.400	.058
	MAIA_Vertrauen	-.005	.312	-.295	.455	.294
Follicular phase	MAIA_Bemerken	.161	-.111	-.196	-.252	.061
	MAIA_NichtAblenken	.048	.188	-.158	.190	-.282
	MAIA_KeineSorgenMachen	-.123	.062	-.081	.209	-.272
	MAIA_AufmerksamkeitsSteuerung	-.056	.357	-.099	.295	.289
	MAIA_EmotionalesGewahrsein	.035	.460*	-.188	.510*	.227
	MAIA_Selbstregulation	-.011	-.025	-.186	-.126	-.285
	MAIA_AufLeibHören	.290	.029	.030	.006	-.007
	MAIA_Vertrauen	.242	.345	-.283	.396	.276
Ovulation	MAIA_Bemerken	.194	.143	-.149	.089	.169
	MAIA_NichtAblenken	.028	.285	-.194	.330	-.173
	MAIA_KeineSorgenMachen	.307	.467*	-.293	.468*	-.065
	MAIA_AufmerksamkeitsSteuerung	.064	.319	-.205	.453	.366
	MAIA_EmotionalesGewahrsein	.031	.206	-.132	.200	.394
	MAIA_Selbstregulation	.179	-.030	-.010	-.235	-.301
	MAIA_AufLeibHören	-.152	-.041	-.047	.028	-.228
	MAIA_Vertrauen	.155	.338	-.289	.271	.151

*Correlation is significant at the 0.05 level (1-tailed)

**Correlation is significant at the 0.01 level (1-tailed)

Table 3.4: Correlations Schandry-scores with Depression-Anxiety-Stress-Scale (DASS) (*N*=14)

		DASS_Depression	DASS_Anxiety	DASS_Stress
Lutealphase	Schandry	.004	-.087	-.077
Follicular phase	Schandry	-.131	-.198	-.122
Ovulation	Schandry	-.365	-.221	-.235

Regarding interoceptive sensitivity, there was a highly significant correlation between the MAIA-II subcategory “Bemerken” and the IFIS subcategories “cardiorespiratory fitness” in the luteal phase $r(14)= .639$, $p= .007$ (s. Figure 4.1) and in the subcategory “speed/agility” $r(14)=.631$ $p= .008$,also in the luteal phase (s. Figure 4.2). There was also a significant correlation between the MAIA-II subcategory “emotionales Gewahrsein” and the IFIS subcategory “speed/agility” in the luteal phase $r(14)= .541$, $p= .023$ (s. Figure 4.3). Furthermore, we found isolated significant correlations in the follicular phase between the MAIA-II subcategory “emotionales Gewahrsein” and the IFIS subgroups “cardiorespiratory fitness” $r(14)= .460$, $p= .049$ (s. Figure 4.4) and “speed/agility” $r(14)= .510$, $p= .031$ (s. Figure 4.5) as well as in the ovulation between the MAIA-II-Subcategory ”keine Sorgen machen” and the IFIS subcategories “ cardiorespiratory fitness” $r(14)= .467$,$p= .046$ (s. Figure 4.6) and “speed/agility” $r(14)= .468$, $p= .046$ (s. Figure 4.7) (s. Table 3.3).

Looking further, we could not observe any significant correlation between the Schandry scores and the Depression-Anxiety-Stress-Scale (DASS) (s. Table 3.4)

Three significant correlations between the Schandry-confidence-task and the DASS were found. The first being between the Schandry confidence and DASS subcategory “depression” in the follicular phase $r(14)= -.566$, $p= .017$ (s. Figure 7.1), the second in for the same categories in the ovulation $r(14)= -.536$, $p= .024$ (s. Figure 7.2) and the third between the Schandry-confidence-task and the DASS subcategory “Stress” in the follicular phase $r(14)= -.477$, $p= .042$ (s. Figure 7.3) (s. Table 3.5).

Moving on, we look at the results from the correlational analyses between the MAIA-II-Questionnaire and the Depression-Anxiety-Stress-Scales, short DASS.

We observed a significant correlation in the luteal phase between the DASS subcategory “depression” and the MAIA-II subcategory “Bemerken” $r(14)= -.590$, $p = .013$ (s. Figure 5.1), “emotionales Gewahrsein” $r(14)= -.465$,$p= .047$ (s. Figure 5.2) and “auf Leib hören” $r(14)= -.496$, $p= .036$ (s. Figure 5.3). Also, a significant correlation in the follicular phase between the DASS subcategory “depression” and the MAIA-II subcategory “nicht ablenken” $r(14)= -.655$, $p= .006$ (s. Figure 5.4) was found. Furthermore a

Table 3.5: Correlations Schandry-Confidence with Depression-Anxiety-Stress-Scale (DASS) (*N*=14)

		DASS_Depression	DASS_Anxiety	DASS_Stress
Luteal phase	Schandry Confidence	-.362	.000	-.367
Follicular phase	Schandry Confidence	-.566*	-.208	-.477*
Ovulation	Schandry Confidence	-.536*	-.072	-.418

*Correlation is significant at the 0.05 level (1-tailed)

Table 3.6: Correlations MAIA-II with Depression-Anxiety-Stress-Scale (DASS) (*N*=14)

Menstruationphase	Variable	DASS_Depression	DASS_Anxiety	DASS_Stress
	MAIA_Bemerken	-.590*	-.157	-.340
	MAIA_NichtAblenken	-.351	-.203	.028
	MAIA_KeineSorgenMachen	-.085	-.268	.071
	MAIA_AufmerksamkeitsSteuerung	-.015	-.101	-.057
	MAIA_EmotionalesGewahrsein	-.465*	-.571*	-.626**
	MAIA_Selbstregulation	.041	.286	.133
	MAIA_AufLeibHören	-.496*	-.460*	-.531*
Luteal phase	MAIA_Vertrauen	-.201	-.510*	-.216
	MAIA_Bemerken	.114	.063	.323
	MAIA_NichtAblenken	-.655**	-.304	-.245
	MAIA_KeineSorgenMachen	-.423	-.258	-.234
	MAIA_AufmerksamkeitsSteuerung	.291	-.122	.116
	MAIA_EmotionalesGewahrsein	-.322	-.460*	-.267
	MAIA_Selbstregulation	-.113	.181	.050
	MAIA_AufLeibHören	-.124	-.034	.052
Follicular phase	MAIA_Vertrauen	-.194	-.353	-.036
	MAIA_Bemerken	-.129	-.154	.056
	MAIA_NichtAblenken	-.491*	-.374	-.224
	MAIA_KeineSorgenMachen	-.313	-.191	.086
	MAIA_AufmerksamkeitsSteuerung	-.397	-.306	-.320
	MAIA_EmotionalesGewahrsein	-.073	-.476*	-.031
	MAIA_Selbstregulation	-.136	.145	.164
	MAIA_AufLeibHören	-.434	-.280	-.379
Ovulation	MAIA_Vertrauen	-.260	-.308	-.019

*Correlation is significant at the 0.05 level (1-tailed)

**Correlation is significant at the 0.01 level (1-tailed)

significant correlation between the DASS subcategory “depression” and the MAIA-II subcategory “nicht ablenken” during the ovulation $r(14)= -.491$, $p = .037$ (s. Figure 5.5) was found. Most correlations between the MAIA-II-subcategories and the DASS subcategory “depression” occur in the luteal phase (s. Table 3.6).

Discussion

This study was motivated by the well-known topic of interoception, as well as pain perception and how they change throughout the menstrual cycle. To investigate the interdependence of objective, subjective, and awareness measures of interoception, we used heartbeat-tracking which issues two objective tests of interoceptive accuracy (e.g., Schandry task), as well as subjective confidence judgements of performance and pain inducing tests.

Interoceptive accuracy, interoceptive sensibility & interoceptive awareness

First off, the most notable outcome is the significant effect in the Schandry task where the interoceptive accuracy score was found at the lowest during the luteal phase. “Interoceptive accuracy” is progressively used to touch on interoceptive behavioral performance (Ceunen et al., 2013) and was measured here by the heartbeat-tracking-task.

The other important result in our data is the marginally significant effect in the heat pain threshold where the pain sensitivity was also found at the lowest levels during the luteal

phase. One possible interpretation for this outcome could be that of a link to it by the female sex hormones progesterone and oestrogen, which have been linked to a lower pain sensitivity, in females' luteal phase as their level are at their peak (Martin MD, 2009; Schertzinger et al., 2017). Explaining also the higher vulnerability found during the ovulation phase, as these hormonal levels are lower. According to a study about luteal analgesia, these high progesterone levels post-ovulation are related to a decrease in the affective element of the pain experience, together with a detachment between pain intensity and unpleasantness. This dissociation is associated with a functional connectivity between the inferior frontal gyrus (IFG) and the amygdala (Vincent et al., 2018). Furthermore, allopregnanolone, progesterone's primary metabolite, has been demonstrated to upregulate the GABAergic system, resulting in an increase in neuronal excitability inhibition (Guennoun et al., 2015; Ikarashi et al. 2020). According to Wang et al. (1996), although there was no assessment of allopregnanolone levels in our study, they are connected with progesterone levels, particularly in the luteal phase. For interoceptive accuracy this could result in overall lower accuracy scores since perception is inhibited by the GABAergic system. For interoceptive sensibility this means that since there is an overall lower perception, the self-perception of being focused on the bodily sensations may be lower as well, but still only in the luteal phase.

We also confirm the lack of a significant effect concerning the cold pain threshold, similar to Kowalczyk et al. (2006) findings, which refers to inconsistencies in pain perception compared to the heat pain threshold. Nonetheless, we noticed that the highest thresholds are demonstrated during the menstruation phase, which is the same as the results of Veith et al. (1984) and Hapidou & de Catanzaro (1988). Overall, this means that the participants were less sensitive during the follicular phase. These inconsistencies in pain thresholds for the different tests can be referred to as the change in women's ability to experience pain during the menstrual phases (Goolkasian, 1980).

Finally, there is no significant effect concerning the interoceptive awareness in terms of the confidence ratings from the Schandry task. Nevertheless, one study found that people with panic disorders or any sort of anxiety disorder have a better cardiac awareness than people with depression disorders. There appears to be a trend toward better interoceptive awareness among individuals with clinical anxiety disorders across clinical groups (Ehlers & Breuer, 1992, Zoellner & Craske, 1999, Pollatos et al., 2009, Dunn et al., 2010, Stevens et al., 2011). Moreover, mild to moderate depression are characterized by a reduced capacity to accurately identify heartbeats, meaning they are negatively correlated and revealing that there is an improvement of interoceptive awareness when participants suffer from severe depression (Dunn et al., 2007; Pollatos et al., 2009). The depression-anxiety-stress-scale (DASS) referred to three significant effects in relation with the Schandry-confidence-task. The lower the score of the depression-anxiety-stress-scale the higher the interoceptive awareness score is. People who are anxious have a higher proclivity for detecting internal body changes, which then becomes the basis of a misinterpreted or mislabelled signal (Garfinkel & Critchley, 2013). However, by Tyrer (1973, 1976), the general pattern of our findings suggests that heightened cardiac awareness is linked to physical forms of anxiety rather than high anxiety per se., thus, being a possible explanation for the significant correlations found between the DASS and the MAIA-II-Questionnaire. Variables like physical activity have also been proven to relate to interoceptive accuracy (Zoellner & Craske, 1999), and therefore explaining the significant correlations found between the IFIS and the MAIA-II-Questionnaire. Another study also revealed that cognitive-behavioural therapy (CBT) influenced a significant decrease in depressive symptoms and a rise in mindfulness and interoceptive abilities tested on a sample of depressive participants (Karanassions et al., 2021).

In summary, the different dimensions of interoception are all affected differently. Alas, so far practically no research has been conducted on this matter. Also, none of the three measured interoceptive concepts show correlations consistently. Yet interoceptive awareness and interoceptive sensibility are more regularly

negatively correlated with depression, being consistent with the found literature (Pollatos et al., 2009). These correlations occur in all cycle phases but interoceptive sensibility mostly in the luteal phase. As for fitness, there are only positive correlations which appear isolated throughout the concepts and cycle phases. For interoceptive accuracy these are in the subgroup of muscular strength, for interoceptive sensibility they are in the subgroups "cardiorespiratory fitness" and "speed/agility", and no positive or negative correlations could be found for the interoceptive awareness.
All in all, fitness generally influences the three interoceptive concepts positively while depression influences them negatively throughout the menstrual cycle. Therefore, our third hypothesis is accepted. As there is only limited research including the different questionnaires (DASS, IFIS & MAIA-II), only minimal explanations could be found for our different results.

Pain perception

For our cold pressor test, the data observed shows no significant effect at all to do with the three menstrual phases, except for the "ceiling effect". Still, we noticed that, just like Teepker et al. (2010), the threshold for the cold-pressor-pain is the highest during the luteal phase and the lowest during the follicular phase, which could also relate to the luteal analgesia, explained by Vincent et al. (2018). In general, we noticed a marginal significance between the three menstrual cycle phases, confirming cycle variations in sensitivity of pain perception. The study of Hellström & Lundberg (2000), using the cold-pressor-test, also revealed that women are less sensitive during the second part of the cycle, e. g. luteal phase, than during the follicular phase due to high oestrogen levels. Interesting to know for these results as well is that progestin-only contraceptive users can have a higher pain tolerance than participants receiving a combined hormonal contraceptive (Máximo et al., 2015). In spite of that, as only 3 participants with a hormonal contraceptive are included in our study, this doesn't have an all-too striking effect on our outcome. Our outcome suggests that hormonal levels could have an important impact on pain sensitivity during the menstrual cycle. Nonetheless, some studies indicate that there is no correlation between hormone levels and pain sensitivity. Other well-controlled studies have shown that the menstrual cycle does not have any effect on the perception of pain in healthy and pain-free women (Granot et al., 2001; Sherman et al., 2005; Kowalczyk et al., 2006; Klatzkin et al., 2010; Teepker et al., 2010; Vincent et al., 2011). Analysing the substantial fluctuations in reproductive hormones throughout the menstrual cycle and considering that animal studies indicated an influence of progesterone and oestradiol on the pain response, it can be astonishing that there is a lack of consistent effects of the menstrual phase on pain sensitivity (Fillingim and Ness, 2000; Craft et al., 2004).

Limitations

The results of our study are limited by the case study sample size, which consists of 14 female participants, who are mostly psychology students. Furthermore, underaged girls, as well as older women than 23 were not taken into consideration. This means the sample is not representative of the population.
Concerning the testing, the German version of the IFIS questionnaire was only validated for adolescents, but since our participants were all between the ages of 18 and 23, we decided to still use it in the sessions. To collect data about the perception of pain we used exclusively thermal stimulation. However, previous research (Sherman & LeResche, 2006) showed that the impact of menstrual cycle phases on pain thresholds is different for different types of stimulation, therefore including e.g., electrical or pressure stimulation would have helped validate our study. In the Cold-Pressor-Test, the water temperatures were not always precise but fluctuated from about 4 to 6 degrees Celsius and since the water did not circulate the conditions under which this testing took place did not allow accurate measurements in the Cold-Pressor-Test.
In further research, contraception should be included as a group factor since the hormonal level seems to be responsible for changes in pain perception and interoception, so it makes a difference whether the participants use

hormonal contraceptives or not. In this context, an addition to the methods of the study could be measuring gonadal hormone levels, because it cannot be assumed that the hormonal milieu of one menstrual cycle is identical to another and to better determine the dates for the testing according to the menstrual cycle phases. In addition, measuring more than one cycle would improve the quality of the data.

Summary & Conclusion

To summarize, the review of the literature suggests a variety of results, implying that if there is a correlation between gonadal hormone levels and pain sensitivity, it's not a simple linear one. Alternative explanations of our findings, mainly the higher thresholds during the luteal phase, not relating to menstrual cycle effects probably exist, such as different types of brain responses. In fact, a few of the already mentioned studies present that even when behavioral pain responses did not change, there are alterations in brain activity patterns in areas implicated in cognitive pain regulation in association with hormonal changes in the menstrual cycle. The brain areas affected were linked to cognitive and motor function rather than pain perception, suggesting that cognitive pain modulation and physical awareness systems are vulnerable to menstrual cycle impacts. (Choi et al., 2006; de Leeuw et al., 2006; Veldhuijzen et al., 2013; Lacovides et al., 2015). Unfortunately, in our study these particular phenomena weren't analysed, otherwise this could be a possible explanation for the moderate link between hormonal levels and variations in pain sensitivity and brain activation in response to painful stimuli during the menstrual cycle (Veldhuijzen et al., 2013). As suggested by Craft (2007), gonadal hormones can affect some but not all forms of pain. Nevertheless, as these analyses were not assessed in the current study, all these interpretations remain partially speculative and other factors that may influence pain perception during the menstrual cycle cannot be excluded.

To conclude, our research provided great new insights of the influence of the menstrual cycle on interoceptive accuracy, sensibility and awareness, as well as pain perception, such as the overall lower pain sensitivity in women during the luteal phase. The demonstrated inconsistencies, along with the described methodological problems, make it difficult to draw conclusions about the relation of experimental pain and the role of pain sensitivity in reference to the menstrual cycle. These inconsistencies might be explained by the small participant sample we provided in this study, compared to larger and broader samples in other studies. In spite of that, with this analysis we still were able to confirm our first and second hypotheses, which stated that there are different levels of sensitivity in pain perception as well as a difference in interoception between the three menstrual cycle phases. All in all, fitness generally influences the three interoceptive concepts positively while depression influences them negatively throughout the menstrual cycle. Therefore, also our third hypothesis is accepted, like we expected. These findings lead our study to successful research and will hopefully be the first of many concerning the investigation of menstrual cycle impacts on interoception.

References

Barrett, L. F., Quigley, K. S., Bliss-Moreau, E., & Aronson, K. R. (2004). Interoceptive Sensitivity and Self-Reports of Emotional Experience. *Journal of Personality and Social Psychology*, *87*(5), 684–697.

Buggio, L., Barbara, G., Facchin, F., Ghezzi, L., Dridi, D., Vercellini, P. (2021). The influence of hormonal contraception on depression and female sexuality: a narrative review of the literature. *Gynecological Endocrinology*, 1–9.

Cameron, O. G. (2001). Interoception: The Inside Story—A Model for Psychosomatic Processes. *Psychosomatic Medicine*, *63*(5), 697–710.

Ceunen, E., Van Diest, I., W. S. Vlaeyen, J. (2013). Accuracy and awareness of perception: Related, yet distinct (commentary on Herbert et al., 2012), *92*(2), 426-427.

Choi, J. C., Park, S. K., Kim, Y. H., Shin, Y. W., Kwon, J. S., Kim, J. S., Kim, J. W., Kim, S. Y., Lee, S. G., Lee, M. S. (2006). Different brain activation patterns to pain and pain-related unpleasantness during

the menstrual cycle. *Anesthesiology 105*(1), 120–127.

Christensen, J. F., Gaigg, S. B., Calvo-Merino, B. (2018). I can feel my heartbeat: Dancers have increased interoceptive accuracy, *Psychophysiology, 55*(4).

Craft, L. L., Perna, F. M. (2004). The Benefits of Exercise for the Clinically Depressed. *Primary care companion to the Journal of clinical psychiatry*, *6*(3), 104–111.

Craft, R. M. (2007). Modulation of pain by estrogens. *Pain, 132*(1), S3–S12.

Craig, A.D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews, 3*(8), 655–666.

Dunn, B. D., Dalgleish, T., Ogilvie, A. D., Lawrence, A. D. (2007). Heartbeat perception in depression. *Behaviour Research and Therapy, 45*(8), 1921-1930.

Dunn B. D., Stefanovitch I., Evans D., Oliver C., Hawkins A., Dalgleish T. (2010). Can you feel the beat? Interoceptive awareness is an interactive function of anxiety- and depression-specific symptom dimensions, *Behaviour Research and Therapy, 48*(11), 1133-1138.

Ehlers A., Breuer P. (1992). Increased cardiac awareness in panic disorder. *J. Abnorm. Psychol.*, *101*(3), 371-82.

Ehlers A., Breuer P., Dohn D., Fiegenbaum W. (1995). Heartbeat perception and panic disorder: possible explanations for discrepant findings. *Behaviour Research and Ther*apy, *33*(1), 69-76.

Eisprungrechner — Jetzt exakter Eisprung berechnen! (O. J.). Mummylator.

Fillingim, R. B., Maixner, W. (1995). Gender differences in the responses to noxious stimuli. *Pain Forum, 4(*4), 209-221.

Fillingim, R. B., Ness, T. J. (2000). Sex-related hormonal influences on pain and analgesic responses. *Neurosci. Biobehav. Rev., 24*(4), 485-501.

Fliege, H., Rose, M., Arck, P., Levenstein, S., Klapp, B. F. (2009). Perceived Stress Questionnaire (PSQ), Leibniz-Zentrum für Psychologische Information und Dokumentation (ZPID) (Hrsg.), Elektronisches Testarchiv. Trier: ZPID.

Frank, R. (Hrsg.). (2011). *Therapieziel Wohlbefinden*. Springer Berlin Heidelberg.

Garfinkel, S. N. (2015). Knowing your own heart: Distinguishing interoceptive accuracy from interoceptive awareness. *Biological Psychology, 104*, 65-74.

Goolkasian, P. (1980). Cyclic changes in pain perception: an ROC analysis. *Perception & Psychophysics, 27*(6), 499-504.

Goolkasian, P. (1983). An ROC analysis of pain reactions in dysmenorrheic and non dysmenorrheic women. *Perception & Psychophysics, 34*(4), 381-6.

Granot, M., Yarnitsky, D., Itskovitz-Eldor, J., Granovsky, Y., Peer, E., Zimmer, E. Z. (2001). Pain perception in women with dysmenorrhea. *Obstet. Gynecol., 98*(3), 407–411.

Guennoun, R., Labombarda, F., Gonzalez Deniselle, M. C., Liere, P., De Nicola, A. F., Schumacher, M. (2015). Progesterone and allopregnanolone in the central nervous system: Response to injury and implication for neuroprotection, *The Journal of Steroid Biochemistry and Molecular Biology*, *146*, 48-61.

Hapidou, E. G., de Catanzaro, D. (1988). Sensitivity to cold pressor pain in dysmenorrheic and non-dysmenorrheic women as a function of menstrual cycle phase. *Pain, 34*(3), 277–283.

Hellström, B., Lundberg, U. (2000). Pain perception to the cold pressor test during the menstrual cycle in relation to estrogen levels and a comparison with men. *Integrative Physiological and Behavioural Science, 35***,** 132–141.

Ikarashi, K., Sato, D., Iguchi, K., Baba, Y., Yamashiro, K. (2020). Menstrual Cycle Modulates Motor Learning and Memory Consolidation in Humans. *Brain Sciences*. *10*(10), 696.

Karanassios, G., Schultchen, D., Möhrle, M., Berberich, G., & Pollatos, O. (2021). The Effects of a Standardized Cognitive-Behavioural Therapy and an Additional Mindfulness-Based Training on Interoceptive Abilities in a Depressed Cohort. *Brain Sciences*, *11*(10), 1355.

Klatzkin, R. R., Mechlin, B., Girdler, S. S. (2010). Menstrual cycle phase does not

influence gender differences in experimental pain sensitivity. *Eur. J. Pain*, *4*(1), 77-82.

Kowalczyk, W. J., Evans, S. M., Bisaga, A. M., Sullivan, M. A., Comer, S. D. (2006). Sex differences and hormonal influences on response to cold pressor pain in humans. *Journal of Pain, 7*(3), 151–160.

Lacovides, S., Avidon, I., Baker, F. C. (2015). Does pain vary across the menstrual cycle? *European Journal of Pain*.

de Leeuw, R., Albuquerque, R. J., Andersen, A. H., Carlson, C. R. (2006). Influence of estrogen on brain activation during stimulation with painful heat. *J. Oral Maxillofac. Surg., 64*, 158–166.

Mahr, E. (1985). Berliner Fragebogen zum Erleben der Menstruation (FEM), Beltz Verlag Weinheim und Basel

Martin MD, V. T. (2009). Ovarian hormones and pain response: A Review of Clinical and Basic Science Studies. *Gender Medicine, 6*(2), 168-192.

Máximo, M. M., Silva, P. S., Vieira, C. S., Gonçalvez, T. M., Rosa-E-Silva, J. C., Candido Dos-Reis, F. J., Nogueira, A. A., Poli-Neto, O. B. (2015). Low-dose progestin-releasing contraceptives are 440 associated with a higher pain threshold in healthy women. *Fertility and Sterility, 104*(5), 1182-1189.

Mehling, W. E. (2012). Multidimensional Assessment of Interoceptive Awareness Version 2 (MAIA-2).

Mehling, W. E., Chesney, M. A., Metzler, T. J., Goldstein, L. A., Maguen, S., Geronimo, C., Agcaoili, G., Barnes, D. E., Hlavin, J. A., & Neylan, T. C. (2017). A 12-week integrative exercise program improves self-reported mindfulness and interoceptive awareness in war veterans with posttraumatic stress symptoms. *Journal of Clinical Psychology*, *74*(4), 554–565.

Migliorini, R. A. (2017). The Structure and Function of Interoceptive Brain Regions in Adolescent Substance Users. *UC San Diego*.

Natale, V., Albertazzi, P. (2006). Mood swings across the menstrual cycle: a comparison between oral contraceptive users and non-users. *Biological Rhythm Research*, *37*(6), 489–495.

Nilges, P., Essau, C. (2021). *DASS. Depressions-Angst-Stress-Skalen — Deutschsprachige Kurzfassung*.

Ortega, F. B., Ruiz, J. R., España-Romero, V., Vicente-Rodriguez, G., Martínez-Gómez, D., Manios, Y., Béghin, L., Molnar, D., Widhalm, K., Moreno, L. A., Sjöström, M., Castillo, M. J., on behalf of the HELENA study group (2011). International Fitness Scale (IFIS), *International Journal of Epidemiology, 40*(3), 701–711.

Parlee, M. B. (1982). Changes in Moods and Activation Levels During the Menstrual Cycle in Experimentally Naive Subjects. *Psychology of Women Quarterly*, *7*(2), 119–131.

Pereira, D. de A., Correia Júnior, J. L., Carvas Junior, N., & Freitas-Dias, R. de. (2020). Reliability of questionnaire The International Fitness Scale: A systematic review and meta-analysis. *Einstein (Sao Paulo, Brazil)*, *18*, eRW5232.

Pollatos, O., Traut-Mattausch, E., Schandry, R. (2009). Differential effects of anxiety and depression on interoceptive accuracy. *Depression and Anxiety*, *26*(2), 167–173.

Procacci, P., Corte, M. D., Zoppi, M., Maresca, M. (1974). Rhythmic changes of the cutaneous pain threshold in man. A general review. *Chronobiologia*, *1*(1), 77-96.

Riley 3rd, J. L., Robinson, M. E., Wise, E. A., Price, D. (1999). A meta-analytic review of pain perception across the menstrual cycle. *Pain, 81*(3), 225-235.

Saupe, R. (1987). *Berliner Fragebogen zum Erleben der Menstruation (FEM): Entwicklung und erste Anwendung*. Huber.

Schandry, R. (1981). Heart beat perception and emotional experience. *Psychophysiology, 18*(4), 483-488.

Schertzinger, M.,Wesson-Sides, K., Parkitny, L., Younger, J. (2017). Daily Fluctuations of Progesterone and Testosterone are Associated with Fibromyalgia Pain Severity. *The Journal of Pain*, *19*(4), P410-417.

Sherman, J. J., LeResche, L., Mancl, L. A., Huggins, K., Sage, J. C., Dworkin, S. F. (2005). Cyclic effects on experimental

pain response in women with temporomandibular disorders. *J. Orofac. Pain., 19*, 133–143.

Sherman, J. J., LeResche, L. (2006). Does experimental pain response vary across the menstrual cycle? A methodological review. *American Journal of Physiology, 291*(2), R245-R256.

Silberstein, S. D., Merriam, G. R. (2000). Physiology of the Menstrual Cycle. *Cephalalgia*, *20*(3), 148- 154.

Stening, K., Eriksson, O., Wahren, L., Berg, G., Hammar, M., & Blomqvist, A. (2007). Pain sensations to the cold pressor test in normally menstruating women: Comparison with men and relation to menstrual phase and serum sex steroid levels. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, *293*(4), R1711–R1716.

Stevens S., Gerlach A. L., Cludius B., Silkens A., Craske M. G., Hermann C. (2011). Heartbeat perception in social anxiety before and during speech anticipation, *Behaviour Research and Therapy, 49*(2), 138-143.

Sutar, A., Paldhikar, S., Shikalgar, N., Ghodey, S. (2016). "Effect of aerobic exercises on primary dysmenorrhoea in college students." *IOSR Journal of Nursing and Health Science*, *5*(5), 20–24.

Teepker, M., Peters, M., Vedder, H., Schepelmann, K., Lautenbacher, S. (2010). Menstrual Variation in Experimental Pain: Correlation with Gonadal Hormones. *Neuropsychobiology, 61*(3), 131–140.

de Tommaso, M. (2011). Pain Perception during Menstrual Cycle. *Curr Pain Headache Reports, 15*, 400–406.

Tyrer, R. J. (1973). Relevance of bodily feelings in emotion. *The Lancet*, *301*(7809), 915-916.

Tyrer, P. J. (1976). The role of bodily feelings in anxiety. London: *Oxford University Press*.

Vaitl, D. (1996). Interoception. *Biological Psychology*, *42*(1–2), 1–27.

Veith, J. L., Anderson, J., Slade, S. A., Thompson, P., Laugel, G. R., Getzlaf, S. (1984). Plasma beta-endorphin, pain thresholds and anxiety levels across the human menstrual cycle. *Physiology & Behavior, 32*(1), 31–34.

Veldhuijzen, D. S., Keaser, M. L., Traub, D. S., Zhuo, J., Gullapalli, R. P., Greenspan, J. D. (2013). The role of circulating sex hormones in menstrual cycle-dependent modulation of pain-related brain activation, *Pain, 154*(4), 548–559.

Vincent, K., Warnaby, C., Stagg, C. J., Moore, J., Kennedy, S., Tracey, I. (2011). Dysmenorrhoea is associated with central changes in otherwise healthy women. *Pain, 152*(9), 1966–1975.

Vincent, K., Stagg, C. J., Warnaby, C. E., Moore, J., Kennedy, S., Tracey, I. (2018). "Luteal Analgesia": Progesterone Dissociates Pain Intensity and Unpleasantness by Influencing Emotion Regulation Networks. *Front Endocrinol* (Lausanne), *23*(9), 413.

Wang, M., Seippel, L., Purdy, R.H., Backstrom, T. (1996). Relationship between Symptom Severity and Steroid Variation in Women with Premenstrual Syndrome: Study on Serum Pregnenolone, Pregnenolone Sulfate, 5 Alpha-Pregnane-3,20-Dione and 3 Alpha-Hydroxy-5 Alpha-Pregnan-20-One. *J. Clin. Endocrinol. Metab*., *81*(3), 1076–1082.

Zoellner L. A., Craske M. G. (1999). Interoceptive accuracy and panic. *Behaviour Research Ther*apy, *37*(12), 1141-58.

From self-concept to study choice

Claire Gend, Ilirjana Havani, Mascha Hilgert, Laura Küpper, Cassandra Origer, Laure Remy

Supervisors: Dr. Ineke Pit-Ten Cate, Dr. Mireille Krischler

For several decades, self-concept has been considered as an important hierarchical construct which has an essential influence on personal thinking, behaviour and decision-making in all areas of life. Specifically for the structure of academic self-concept, the Marsh-Shavelson-Model plays an important role and has a huge impact on psychological research (Marsh, 1990). In accordance with this model, the academic self-concept is divided into a math and verbal self-concept containing domain specific subjects. Based on the Internal-/External-Frame of Reference Model, the development of (explicit) self-concept can be explained by social (interindividual) and dimensional (intraindividual) comparisons. Furthermore, scholastic achievement such as school grades affect the self-concept (Möller et al. 2020). In comparison the implicit self-concept affects personal behaviour due to unconscious motives and attitudes. The present study examines the academic implicit and explicit self-concept as well as school grades with the aim of finding out how these factors predict the decision of study choice either in the domain of natural sciences or humanities. Furthermore, the relation between both self-concepts and school grades is investigated and whether school grades directly influence the study choice or if self-concept acts like a mediator between both variables. Gender and socioeconomic status are treated as control variables. To collect our data, 97 high-school or first year university students performed an Implicit Association Test and answered an adapted version of the “Academic Self-description questionnaire” (Marsh, 1990). The main results of our study concluded that in combination with the implicit self-concept only the explicit self-concept has an influence on study choice. Furthermore, a significant association between school grades and explicit self-concept is found, whereupon there is no relation between school grades and implicit self-concept. Lastly, self-concept does not function as a mediator between school grades and study choice, since study choice can be directly predicted by school grades.

1. Introduction

“What are your plans now, since you have finished high school?” Every year, thousands of students complete their school careers after usually 12 or 13 years and are faced with the question, "What's next?". The “Abitur” is the highest school-leaving qualification in Luxembourg and Germany and opens a lot of doors for the graduates to shape their individual future. Since career orientation begins particularly in the last two years of secondary school, students usually already deal with this question intensively during this time. Representatives from universities visit schools and present their courses of study, a number of apprenticeships are presented at school fairs, and (international) organisations advertise au pair stays or work & travel-offers abroad. The opportunities nowadays seem endless, and it is by no means easy to decide on a direction and to stand behind it completely. If a school leaver ultimately decides to begin his or her studies, the specific choice of study is not necessarily reduced: a variety of fields of study offer the entry into any domains which are becoming ever more diversified. According to that, how do you finally decide on a course of study? First and foremost, the choice of study should ideally go hand in hand with one’s personal interests and match with their desires. Nevertheless, school grades play an essential role in whether you can even be admitted to a degree program at all. In addition to these obvious influential factors, subtle factors such as parents’ privily

expectations or gender-based stereotypes may be important aspects for one's decision as well.

The aim of our study is to find out which variables influence the individual study choice. Therefore, we investigate the relationship between school grades, academic self-concepts and study choice in consideration of gender effects and socioeconomic status. Nevertheless, the theoretical findings available on these topics will be presented in more detail below.

2. Theoretical Background

The "APA dictionary of psychology" defines self-concept as an individual's description and evaluation of oneself connected to the personal picture of their own identity. This self-concept contains a self-image including both psychological and physical characteristics for example skills and social roles (VandenBos & American Psychological Association, 2015). Moreover, the general self-concept can be distinguished in an explicit and an implicit self-concept. Whereas implicit motives are usually not under conscious control and managed automatically, explicit motives are conscious and influence thinking and behaviour directly. The individual is therefore able to express and communicate about his/her explicit motives. Since internal mental representations are usually visible only to a certain degree and difficult to explain by the performer, an Implicit Association Test can be used to measure attitudes towards specific subjects (Hofmann et al., 2005). Explicit motives can be measured more easily by self-report-questionnaires, because the performer is aware of them and able to reflect on them properly. Nevertheless, both explicit and implicit self-concept influence behaviour, judgment, decisions, and mood in different situations. Self-concept is an important factor in positive psychology which includes positive beliefs and focuses on the outcome of life for the normal population. In this context, self-concept functions as a predictor and is regarded as a mediator between life decisions, persistence of oneself, and preferable outcome (Möller et al., 2020). The self-concept as such changes throughout the whole lifespan, as well as the personal value of one's self-concept. Throughout a meta-analysis, Möller et al. (2020) reveal this effect between students from elementary school and secondary school. With increasing age, as awareness of individual competences, personal strengths and weaknesses increase, it results in a self-concept more strongly correlating with external factors. Furthermore, Marsh et al. (2015) conclude that different self-concepts are formed in different domains based on e.g. personal experiences or social influences. Research based on the topic of self-concept has grown and improved immensely due to the increase of quality of measurement instruments and research designs over the last 40 years and will be inspected closer in the following paragraphs (Marsh et al., 2015). In essence, the present study will investigate the relationship between implicit and explicit academic math or verbal self-concept in relation to school grades and the prediction of study choice.

The construct of self-concept is nowadays considered as a multifaceted and hierarchical model (Liu et al., 2005). In studies in the 1970's, Shavelson et al. (1976) foreground the hierarchical structure while developing a self-concept model with a first- and second order, dividing the general individual self-concept into an academic self-concept and a non-academic self-concept. Basically, the academic self-concept simply includes specific self-concepts in specific domains. More precisely, self-concepts in each school subject are built and function independently of each other and vary across school subjects in terms of their expression. In comparison, the non-academic self-concept contains specific social, emotional and physical self-concepts. The latter is only mentioned for the sake of completeness, nonetheless, only the academic self-concept has an important relevance in our study and will be focused on supporting and carrying the research questions and hypotheses. In further research in the 1980's (N = 758), Marsh and Shavelson were able to approve the hierarchical order of self-concept, however it was not possible to confirm the structure of the subordination since it is more complex than previously assumed. As a result, they expanded the preceding model and specified the lower ranks resulting in the "Marsh/Shavelson-Model of Academic Self-

Concept" in 1985 and constructed the "Academic Self-Description Questionnaire" (Marsh, 1990). Based on this model, the general academic self-concept is divided into a math academic (including the subjects Mathematics, Physics, Biology and Chemistry) and a verbal academic self-concept (including the subjects English and other languages, History and Geography). Results of Marsh's investigation (1990) show that school subjects within a domain are significantly higher correlated than between the domains.

To get a closer look into the suborders of both verbal and math self–concepts, the development of one's individual general academic self-concept has to be explained beforehand. According to the Internal vs. External Frame Reference Model of Self-Concept, the domain specific academic self-concept is generated through social and dimensional comparisons and dependent on the frame of reference (Marsh et al., 2015). Basically, social comparisons mean interindividual/external comparisons, where the student compares personal performance and achievement to other peers' achievement. On the contrary, dimensional comparisons are based on intraindividual/internal comparisons. Therefore, the student compares personal achievement in one school subject to achievement in other school subjects. Theoretically, temporal comparisons do have an influence on the frame of reference for the general self-concept as well. According to Möller et al. (2020), social and dimensional comparison effects are stronger, whereby temporal comparisons are often of less value specifically for the academic self-concept. Because of that, they will not be put on record in more detail and no items relating to this will be used in the questionnaire of the present study. Achievement specifically defines the skills and competences an individual has obtained (Genesee, 2006), in scholastic terms it is for example indicated in form of school grades, teacher ratings or objective test scores (Trautwein et al., 2006). In this regard, it appears that achievement does play an important role for the academic self-concept and affects the self-concept within as well as between students (Möller et al., 2020). At the same time, it is important to mention that even if the achievement of different students seems to have the same objective characteristics, the social and dimensional comparisons may lead to different subjective perceptions concerning the individual self-concept in the domains (Marsh et al., 2015). The same objective characteristics describe for example the same school grade of two students in Math (social comparison) implicating that the performance of both students in this subject is similar. Even if this is the actual case, it does not mean that both students perceive their mathematical self-concept in the same way. If one of the students acquires significantly better or worse school grades in (e.g.) history than in math (dimensional comparison), this affects the formation of both math and verbal academic self-concept to a great extent. According to that, Möller's et al. (2020) investigation (N = 507,787) reassured a positive correlation between both verbal and math achievement (r = .48).

In a conducted study (N = 117,321), Marsh et al. (2015) test the generalizability of the Internal and External Frame of Reference Model in 13 countries. They assert that achievement in the same domain is positively correlated to the academic self-concept in the corresponding area (r = .534). Hence, high achievement in the mathematical area is related to a higher math self-concept and high achievement in the verbal area is related to a higher verbal self-concept. Coincidentally, the achievement in one domain is negatively correlated to the other domain's self-concept (r = -.235). The results in the conducted meta-analysis of Möller et al. (2020) additionally confirm that verbal and math self-concept are nearly uncorrelated (r = .09). Hence, students rate themselves either as a verbal or math orientated person which may explicitly influence the individual choice of study (Marsh et al., 2015). As mentioned above, self-evaluation and self-rating applies in particular to the (subjective) ratings of teachers. Hereby, the self-rating of personal scholastic competences and achieved performances in many cases coincides with the external subjective perception and assessment of the teacher. At this point, it is interesting to note that school grades and academic self-concept in the corresponding domain reflect a correlation with a medium effect size (b = .50 for first language; b = .60 for math) and are more positively

correlated than academic self-concept and standardised test results, even though the evaluation seems more objective (Möller et al., 2020). Marsh et al. (2005) reasoned that the own and the peers' school grades represent a salient indicator for the source of feedback and depict a helpful target value for social and dimensional comparisons. Students are familiar with how school grades are created, while standardised tests often do not even publish individual scores. It is not transparently obvious to students which concrete individual performances lead to the result (Möller et al., 2020). Thus, school grades can be directly traced back to the domain specific academic self-concept and release motivational effects. However, this phenomenon is accompanied by the students' beliefs that teachers know better about their performance and achievement and are in a better position to assess students' academic competences than they are able to do themselves. Self-reflection and self-assessment are thus restricted to a considerable extent (Marsh et al., 2015).

During this, the implicit academic self-concept plays an essential role. As mentioned in the beginning, the displayed behaviour of an individual is influenced, amongst others, by non-consciously controllable implicit attitudes. This behaviour is mediated by the internal mental representation that is attempted to be adhered to (Greenwald et al., 1998).

In terms of study choice, this compliance includes, for example, maintaining the achievement of good grades in order to be able to start the dream study. In a psychological review, Greenwald and Banaji (1995) reveal different influencing psychological phenomena that explain the development of the implicit self-concept. Approving or disapproving implicit attitudes toward persons, objects, etc. can be attributed for example to the Mere-exposure-effect. The mere perception of something previously judged neutral is eventually evaluated more positively by repeated perception. Students usually perceive their parents' occupation (which is associated with socioeconomic status) daily. Because of that, socioeconomic status may affect the choice of study, meaning more precisely that parents' degrees and occupation have an impact on the chosen domain of study.

Furthermore, internalised and learned (often unconscious) social role expectations and stereotypes play a decisive role here as well. In a study in 2002, Kessels and Hannover show that career choice is significantly related to gender (N = 183). Girls (M = 3.01) are more likely than boys (M = 2.40) to choose a profession in the verbal/humanities domain. On the other hand, boys (M = 3.76) tend to choose a profession in the scientific domain more often than girls (M = 3.56). Consequently, implicit and explicit attitudes can differ, however, there is by no means a conscious awareness of these differences.

2.1 Objectives

Based on the presented theoretical background, we define the objectives of our study. First, it can be stated that self-concept is identified as an important factor for decisions and executed behaviour including educational trajectories (Kessel & Hannover, 2002). Particularly in our study, the decision comprehends the variable "study choice", since less research is available here. More precisely, the purpose is to investigate more fully, whether the personal study choice is influenced by both implicit and explicit academic self-concept. In the present study, academic self-concept is divided into humanistic and natural scientific self-concept, however in terms of school subjects, they are in accordance with the beforementioned verbal and math self-concept. We used English, German, History and Social Sciences as subjects for the self-concept of humanities. To test the self-concept in the natural sciences, Math, Chemistry, Physics and Biology are the considered subjects. Secondly, school grades do in fact affect the self-concept and may display a strong predictor for one's individual study choice. In our study, we will analyse to what extent the implicit and explicit academic self-concept (and whether it is rather humanistic or scientific orientated) are related to school grades. Finally, it should be scrutinised whether school grades are directly related to study choice or if the effect of grades on study choice is mediated by self-concept. Gender and socioeconomic status will be treated as control predictors to

reduce the influence of possible confounding effects.

Consequently, we postulate three main hypotheses:
1. The explicit and implicit self-concept have an impact on the study choice.
2. There is a domain specific positive association between school grades and self-concept.
3. Self-concept mediates between school grades and study choice.

3. Method

To test the hypotheses, an adapted "Implicit Association Test" (Greenwald et al., 1998) and a self-questionnaire based on the "Academic Self-Description Questionnaire" (Marsh, 1990) are conceived. In the following, the study sample is described and further details about the study methods will be explained.

3.1 Participants

In order to design the study, the number of participants necessary to find a possible significant effect in the data has to be defined. This sample size is determined by the "G-Power 3.1" analysis program. We performed an a priori statistical power analysis using linear multiple regression to support our hypotheses. The number of tested predictors is three (school grades, explicit self-concept, implicit self-concept) and the total number of predictors is five (school grades, explicit self-concept, implicit self-concept, gender, socioeconomic status). In a meta-analysis, Möller et al. (2020) reported an average medium effect size for the association between grades and self-concept, whereas Kessels and Hannover (2002) stated that academic self-concept is an important determinant of educational trajectories. Therefore, the effect size is set on 0.15, the statistical power is set on 0.8 and the alpha-level on .05. The analysis ends by giving us the number of participants necessary, $N = 92$.
In our study, 98 participants in total are recruited. Among them, 75.5% ($N = 74$) are female and 23.5% ($N = 23$) male. One participant did not answer this question wherefore no assignment can take place. The ages of the participants range from 16 years to 25 years ($M = 19.86$, $SD = 1.347$). We recruited the participants with the help of social media while sharing the link in Instagram-stories, Facebook-posts or through personal contacts. In order to be able to look at a wide range of different subjects, we tried to reach participants from as many different courses of study as possible. 63,3% ($N = 62$) of our participants have chosen a subject of study in the domain of humanities and 32,7% ($N = 32$) in the domain of natural sciences. All participants are in their final year of secondary school or in their first year at university. This ensures that they are currently considering their choice of study or have recently made a choice. In effect, the respondents have a current connection to our study and can answer the questionnaire without any problems and know their grades well, which is important for the self-questionnaire in the second part of the study. 43.9% ($N = 43$) of the participants graduated or will graduate from school in Luxembourg, 46.9% ($N = 46$) in Germany and 3.1% ($N = 3$) in Belgium. 4.1% ($N = 4$) stated otherwise and 2% gave no answer ($N = 2$). We also determined the highest occupation level of both parents. Of these, 76 parents are skilled and 13 are unskilled.

3.2 Procedure

Our study is an approximately 30 minutes computer-based online study programmed by the "Inquisit"-Software (Version 5.0.14.0) and conducted during a period of four weeks. Subjects can therefore complete the self-questionnaire and the Implicit Association Test from their own laptops/computers. If this option is not available for individual participants, it is also possible to complete the study on one of the laptops provided by the University of Luxembourg. The respective links to the study are shared privately or publicly on our social media accounts. The experiment is offered in three languages (German, French and English) to obtain a sample as heterogeneous as possible. As soon as participants open the link to the experiment, they receive some information about the content

and objectives of the study. It is also transparently explained that it is an anonymous participation. There are also two options: on the one hand, the study can be started (whereby the participants agree to take part in the experiment and consent to the processing of their data). On the other hand, it is possible to stop the study at this point. The anonymity of the test subjects is guaranteed, as no names are requested for data collection. This personal protection is ensured by giving each subject an individual and automatically generated number. If the participant decides to continue with the study, the first step is a reaction-timed test referring to the Implicit Association Test (IAT). The IAT includes various tasks in order to measure the participants' implicit academic self-concept of both scientific and humanistic domains. The second part of the study refers to the measurement of explicit academic self-concept through a questionnaire. In the last step, participants are asked to answer questions regarding their demographic information.

3.3 Measures

To be able to measure the academic self-concept separated as implicit and explicit self-concept, we conduct two different types of tests.
The implicit self-concept reflects implicit attitudes that are usually not under conscious control. Since there is a link between attitudes and executed behaviour (Marsh,1990) we want to measure this without influence, such as reactivity. These attitudes are measured by indirect measures so that the subject is not aware of what the targeted variable is. As mentioned above, the implicit self-concept is measured by an adapted version of the Implicit Association Test (IAT), requesting implicit attitudes towards various concepts, which are the focus of the measurement and the attribution (Greenwald et al., 1998). In our study, this concept reflects the academic self-concept for the domain of natural sciences and the humanistic academic self-concept. Participants were asked to conduct the IAT and associate personal pronouns either self-related or other-related to the category of natural sciences or humanities. The categorisation is based on a given predefined key combination. The IAT contains different blocks consisting of a practice pass and a real pass afterwards. For the first task, the targeted concept shows up in a 1-choice task (e.g. self-related pronouns are to be assigned to natural sciences; other-related pronouns are to be assigned to humanities). In the second task, the attribution appears in a 2-choice task (e.g. self-related pronouns and domain correspondent school subjects are to be assigned to natural sciences; other-related pronouns and domain correspondent school subjects are to be assigned to humanities). Afterwards, the combinations reverse. Participants shall answer as quickly as possible and try to keep their error rate as low as possible. Wherever the categorisation takes less time to perform, it reflects his/her tendency for one of the two academic self-concepts, either towards natural sciences or humanities. For example, if a participant is faster with the association of self-related pronouns and natural sciences than with the combination of self-related pronouns and humanities, he/she is more likely to have an implicit academic self-concept leaning towards the natural sciences. The system automatically deletes all values of the exercise pass. Reaction times between 300ms and 3000ms are recorded. However, reaction times under 300 milliseconds are recoded as 300 milliseconds and reaction times over 3000 milliseconds are recoded as 3000 milliseconds (Greenwald et al., 1998). On the one hand, error values can be corrected. On the other hand, participants who performed the IAT incorrectly are detected and can be excluded from the sample.
The explicit academic self-concept is measured with an adapted version of the "Academic Self-Description Questionnaire", short ASDQ (Marsh, 1990), in either German, English or French. The trilingual version ensures that as many test takers as possible can complete the tests in their mother tongue. The ASDQ is used to record the subjects' self-assessments of their school performance in social (interindividual) and dimensional (intraindividual) comparison. Participants should answer questions concerning their perceived academic achievement in school subjects in comparison to other peers' achievement and their own achievement in other school subjects. The examined school

subjects are German, English, French, History and Social Sciences that are embraced under the domain of humanities, and Math, Physics, Chemistry and Biology that fall in the category of natural sciences. The self-evaluation is based on a seven-point Likert-scale, whereas (1) means “strongly disagree” and (7) “strongly agree”. Finally, questions concerning the personal school grades are added, which are objective performance information. In order to get a more precise overview of the sample, demographic information such as gender, age, mother tongue, country in which the school leaving certificate was completed and socioeconomic status were requested, as well as the participants' study choice. The socioeconomic status is targeted by questions regarding the job (and degree) of the participants' parents. Whether the parents’ education is considered as skilled or unskilled is determined based on the ”International Standard Classification of Occupations” (International Labor Office, 2013), which classifies occupations into a hierarchical classification scheme. Parents whose occupations fell into one of the categories managers, professionals, technicians and associate professionals or clerical support workers were coded as skilled. On the other hand, parents whose occupation fell into one of the categories services and sales workers, skilled agricultural, forestry and fishery workers, craft and related trades workers, plant and machine operators and assemblers or elementary occupations were coded as unskilled.

3.4 Statistical Analysis

The data is analysed using the statistics program "IBM SPSS Statistics 27". In the first place, we conducted descriptive analyses such as the mean, standard deviation, number of participants and frequencies to get a general idea of the study sample. Analyses of a (stepwise) binary logistic regression are used to measure the influence of implicit and explicit self-concept on study choice and to determine whether the effect of grades on study choice is mediated by self-concept or whether grades directly influence study choice. This answers our first and third research question. Furthermore, while using a logistic regression as well, we examine if extraneous variables such as gender and socioeconomic status have a significant effect on the dependent variable, being study choice. To determine the relationship between self-concept and school grades, meaning to answer the second research question, a Pearson-Correlation and a linear regression is used.

4. Results

First of foremost, it is important to mention that all analyses consider the domain of natural sciences as a reference group.

4.1 Self-Concept as a predictor of study choice

As stated in the methods section, the relationship between self-concept and study choice is measured by conducting a binary logistic regression. As indicated in *Table 1*, results show that explicit self-
concept in the domain of humanities and natural sciences functions as a predictor of study choice. From the Chi-square analysis, one can conclude that the full model is statistically significant compared to the constant only model $X^2(4, N = 91) = 25.22$, $p < .001$. The overall prediction success is 75.8%, as well as 86.7% for the domain of humanities and 54.8% for natural sciences. The explicit self-concept in the domain of natural sciences is a highly significant predictor for study choice, which is revealed by the odd-ratios. They indicate that participants with a higher implicit or explicit self-concept in humanities are less likely to choose to study in the field of natural sciences, whereas participants with a higher self-concept in natural sciences are about 2.28 times more likely to choose to study in that field.

Table 1. Self-concept as a predictor of study choice

	Odds ratio	*Wald*	*p*-value
Explicit Self-Concept Humanities	.571	4.18	.04*
Implicit Self-Concept Humanities	.996	3.32	.07
Explicit Self-Concept Natural Sciences	2.28	12.08	.001**
Implicit Self-Concept Natural Sciences	.997	1.76	.18
Constant	.32	.46	.50
Nagelkerke $R^2 = 0.34$			

Furthermore, the implicit self-concept, in both humanities and natural sciences is not a significant predictor of study choice. To summarise, Nagelkerke's R^2 demonstrates that 34% of the variance in study choice can be explained by the indicators of self-concept.

4.2 Domain specific association between school grades and self-concept

Results of a Pearson-Correlation analysis support our hypothesis that there is a domain specific positive association between school grades and self-concept, especially between school grades and the explicit self-concept (*Table 2*). In the domain of humanities, there is a moderate significant correlation between explicit self-concept and grades. The same applies for the domain of natural sciences. However, there does not seem to be a significant correlation between school grades and the implicit self-concept in both humanities and natural sciences. Based on the scoring, a negative correlation coefficient is indicative of a positive association, i.e. lower grades are indicative of better achievement whereas higher ratings of self-concept reflect stronger self-concept.

Table 2. Pearson-Correlation between school grades and implicit/explicit self-concept

	1	2	3	4	5	6
1. Grades Humanities		.387**	-.027	-.512**	.202*	-.008
2. Grades Natural Sciences			.021	.150	-153	-.507**
3. Implicit Self-concept Humanities				.123	-.721**	-.126
4. Explicit Self-Concept Humanities					-.151	-.036
5. Implicit Self-Concept Natural Sciences						.106
6. Explicit Self-Concept Natural Sciences						

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).

Besides the Pearson-Correlation analysis, we conducted a linear regression including school grades as the independent variable and self-concept as the dependent variable of both domains, to see which one predicts the other. Results of the regression analysis indicate that school grades are significant predictors of self-concept in the domains of humanities $F(1, N = 66) = 34.155$, $p < .001$ and natural sciences $F(1, N = 25) = 32.872$, $p < .001$. The R^2 for the domain of humanities indicates that 26.2% of the variance in self-concept can be explained by the indicators of school grades in that specific domain, whereas the R^2 for the domain of natural sciences is 25.7%.

4.3 Self-concept as a mediator between school grades and study choice

Table 3. Predictors of study choice (mediation analysis)

	Odds ratio	*Wald*	*p*-value
Step 1			
School Grades Humanities	2.96	9.13	.003**
School Grades Natural Sciences	.29	15.43	.000**
Constant	.93	.009	.925
Nagelkerke $R^2 = 0.37$			
Step 2			
School Grades Humanities	3.55	6.46	.011**
School Grades Natural Sciences	.33	7.29	.007**
Explicit Self-Concept Humanities	1.24	.34	.56
Implicit Self-Concept Humanities	.994	4.44	.04**
Explicit Self-Concept Natural Sciences	1.498	1.796	.18
Implicit Self-Concept Natural Sciences	.996	1.99	.16
Constant	.30	1.98	.15
Nagelkerke $R^2 = 0.46$			

A stepwise binary logistic regression was conducted to investigate the third research question (*Table* 3). Before looking into the relation between school grades, self- concept and study choice, we included gender and socioeconomic status as extraneous control variables. Cohen's κ was conducted to determine whether there was a meaningful agreement between two raters on socioeconomic status. A strong significant agreement between the two raters' judgements was found, $\kappa = .917$, $p < .001$. Results of the logistic regression analysis show that if we only consider the control variables, the model is significant, $X^2(2, N = 89) = 6.04$, $p < .05$. This is caused by an effect of gender. Gender is significant when it is the only variable applied as a predictor of study choice, $W(1, N = 91) = 5.512$, $p < .05$. However, in combination with the other variables, it does not have a significant influence on study choice, $W(1, N = 91) = 2.820$, $p > .05$. For the socioeconomic status (SES), we differentiated between skilled and unskilled professions of the students' parents and we came to the same conclusion that if combined with the other variables, the SES does not have a significant influence, nor does it significantly

predict study choice, $W(1, N = 89) = .012$, $p > .05$. Also, SES has no significant influence on study choice, when it is the only variable applied besides study choice. Due to the fact that both gender and socioeconomic status are not seen as predictors of study choice, we do not consider these control variables in the following analysis.

In the first step of the binary logistic regression, school grades were entered as predictors of study choice. From the Chi-square analysis, one can conclude that the full model is statistically significant compared to the constant only model $X^2(2, N = 91) = 28.33$, $p < .001$. The overall prediction success is 68.9% (86.4% for humanities, 35.5% for natural sciences). Results in *Table 3* indicate that school grades in both humanities and natural sciences are significant predictors of study choice. Considering the odd-ratio, we can see that for every unit that grades in humanities increase - meaning a worse achievement - the chance that the person will choose to study natural sciences increases with a factor of 2.96, whereas for every unit that a student achieves poor performance in natural sciences and experience a degradation of the grades, the chance that the person will choose to study natural sciences is reduced with a factor of .29. Combined, school grades can explain 37% of the variance in study choice.

In the second step, self-concept is included in the model of the binary logistic regression. Although the Nagelkerke R^2 increases to 46%, the Chi-square analysis indicates that adding these predictors does not significantly improve the model, $X^2(4, N = 91) = 8.43$, $p = .08$. Contrarily, the whole model is significant, $X^2(4, N = 91) = 36.752$, $p < .001$. In other words, indicators of self-concept do not significantly predict study choice after controlling for school grades.

5. Discussion

The aim of the present study was to determine the relationship between implicit and explicit self-concept, school grades and study choice. With our analyses we can confirm, as well as reject our hypotheses.

5.1 Hypothesis 1

Results of a binary logistic regression showed that self-concept has an impact on the choice of study. More specifically, study choice could be significantly predicted by both explicit self-concept in humanities and explicit self-concept in natural sciences. Our first hypothesis is therefore partially rejected because the implicit self-concept in natural science and humanities has no significant impact. This can be explained by the fact that one will be more likely to pursue studies in a field in which they perceive themselves to be successful and in which they are explicitly more interested in. Success in a subject can for example be expressed by good grades or good teacher-ratings. In fact, if one gets good grades in a school subject, the possibility that one will also choose a related career path is higher than for a school subject in which one gets lower grades and as a result the explicit self-concept is strengthened. Because the implicit self-concept contains our unconscious attitudes, it is less visible. The explicit self-concept includes attitudes that are consciously accessible to us and thus directly influence behaviour. For this reason, it is more internalised and has a stronger influence on study choices than the implicit self-concept.

5.2 Hypothesis 2

The Pearson-Correlation analysis revealed a significant relationship between school grades and explicit self-concept. Moreover, using a linear regression shows that school grades are a significant predictor of the self-concept in humanities and natural sciences, which is already described by Möller et al. (2020). This is the case because conscious comparisons of grades are made on a social and dimensional level. Comparing grades with other students or comparing one's own grades can have both positive and negative effects on self-concept. Good grades are usually motivating and therefore have a positive effect on the self-concept, while bad grades tend to be demotivating. However, lower grades can also increase motivation as students try to improve their performance and achieve better results. Thus,

comparing ourselves with a student who has lower grades increases our self-concept and comparing ourselves with a better student has a negative effect on our self-concept. It is important to note that such comparisons always depend on the frame of reference. For example, if you always get good grades in math, but the person or subject you are comparing yourself to is better than you, this will have a different effect on your explicit self-concept than comparing oneself to someone with worse grades. Various psychological theories confirm that people strive for a positive self-image and self-concept. To satisfy this need, they seek consistent information that enhances the self-concept. Grades are beneficial in this respect, as they provide a quick and uncomplicated basis for comparison due to their standardisation in the respective school system. However, no correlation was found between school grades and implicit self-concept. From this we can conclude that our conscious self-concept has an influence on whether we are more likely to get good or bad grades. Our unconscious self-concept, on the other hand, does not play such a big role. This could be explained by the fact that grades are more likely to be used explicitly through the earlier mentioned dimensional and social comparisons, while our implicit self-concept is more likely to encompass our unconscious attitudes. In conclusion, we can partially accept our second hypothesis.

5.3 Hypothesis 3

The results of the stepwise logistic regression reveal that gender and socioeconomic status have no influence on study choice when considered in addition to the other variables. However, the variable gender itself, without looking at the other variables, does have an influence on the choice of study. This could be explained by the gender-specific stereotype, which was also described in the study by Kessel and Hannover (2002). Girls are used to being more interested in the humanities and boys in the natural sciences and are therefore more likely to choose studies in this direction. Nevertheless, the fact that school grades are a stronger predictor for study choice than gender can be explained that in our Western society young people are usually encouraged to pursue individualistic goals which includes personal self-development free from gender-stereotypic expectations. For this reason, grades are thus more salient than gender as they make (apparently) objective statements about performance that could also ensure success at work later on. Nowadays, there is a conscious effort to make sure that both boys and girls can gain an equal insight into different domains, for example at career fairs. In addition, many science-oriented degree programs, universities or companies are trying to make natural science fields more attractive to girls and humanities fields more attractive to boys attempting to break down gender stereotypes and counteract prejudices.
The statistical analysis reveals that if you have better grades in a subject, you are more likely to study in that field later on and vice versa wherefore study choice can be predicted significantly by school grades. Adding indicators of implicit and explicit self-concept to the equation did not significantly improve the model. This indicates that there may be shared variance between grades and self-concept in the prediction of study choice. Therefore, no support for the mediation was found, and consequently, our third hypothesis is rejected. These results show that school grades and self-concept do not have a unique effect on the study choice, as described for example by Möller et al. (2020). In addition to grades, objective tests and teacher ratings are also used to determine the academic self-concept, although these are less decisive for the formation of the self-concept. This can be explained by the fact that grades are the strongest indicators of one's performance during the school career, as they appear most frequently in assessments and therefore stand out. In the life of a competitive performance society, grades in the school system are the simplest tangible and quantified unit for making social and dimensional comparisons. A very good grade, which corresponds to a “1” (best grade in the German school system), is in any case always a “1” and implicates objectively a better achievement than a student receiving a school grade of “2”. How the comparisons affect the self-concept always depends on the frame of reference. As a result,

and as mentioned above, grades are more salient and are more likely to be used for choosing a course of study. Especially in science subjects, which are seen as more difficult and demanding in society, grades are a crucial indicator and reason for choosing a field of study in this domain. Here, the construct of self-efficacy plays a role in whether students assess themselves and their abilities so that they will successfully complete a course of study in the natural sciences. If students already achieve good grades in science subjects at school, this influences their attitude toward their own competences. If we come back to the striving for a positive self-image, students already have a certain guarantee to be able to master the study successfully. More precisely, if they already achieve good grades in this domain during their school years, they will step into the study more self-confident which may have an effect on the achievement during their studies. Thus, grades can be seen as a protector of the self-image.
In addition, many degree courses in Germany require a numerus clausus (NC), which in turn expresses the importance of grades. The school and university system thus requires from the outset that one has the appropriate grades in order to be able to take the desired course of study, wherefore students learn from the very beginning the relevance of school grades.

6. Limitations and Outlook

Even though our study revealed significant results, limitations of this research have to be considered and outlooks for further studies can be made. First, we have more female than male participants in our study sample whereas most of the participants have chosen a humanities degree program and are more likely to have a humanities oriented self-concept. This might be a possible confound because boys probably still tend to prefer to choose a field in the domain of natural sciences and girls in the field of humanities. For these reasons, our study is not necessarily well generalisable, as gender and majors are not equalised. To assure more precise results, future studies should try to recruit the same amount of participants that study humanities and natural sciences. This would allow it to make it possible, to look more closely at gender differences, whether women really are more likely to have humanities oriented self-concept than men and whether there are significant differences to previous years and why this is (still) the case. Besides, the fact that we only have participants from the Western European school system means that our study sample may not be representative for the general population. Since our sample mainly contains students from Luxembourg and Germany, the school systems are very similar and can for example not be well compared to the school system in Asian countries. Different countries have different school systems and sometimes also different concepts of learning. In some cultures, lesson contents for example mostly focus on subjects in scientific domains or are mostly humanities orientated. This may have an impact on the students' grades and self-concept since they are socialised differently from the outset. Future studies could include examining cultural differences more closely to be able to see how much these differences influence the choice of study. Furthermore, it might be interesting to analyse, whether the impact of school grades on study choice can be found cross-cultural as well, to generalise the effect or whether the relevance of grades is lower in other school systems.
Moreover, a longitudinal study can be conducted to examine the relationship between self-concept and grades further. In that way a longer period of time can be analysed and it would be possible to learn more about the development of the academic self-concept. In our study, we do not include changes or developments in grades, but only ask the participants about grades at one point in time. During a longitudinal study, significant changes in individual school grades can be registered that may therefore change the explicit self-concept over time. In this way, it can be scrutinised whether intraindividual temporal comparisons may not play a more decisive role than previously assumed.
In our study we only focused on how grades and self-concept influence the study choice.

Another interesting approach would be to add enjoyment as another variable.
All in all, we can conclude that the academic self-concept does have an influence on study choice. Nevertheless, in combination with school grades, the impact is significantly lower, which shows the immense influence of grades on the choice of study. Based on the results, we would like to give one last critical thought-provoking impulse: although grades certainly say a lot about the school performance, it must always be kept in mind that they can exert enormous pressure on students at the same time. It should therefore be considered that good grades in a subject are probably a good precondition for a field of study in a similar domain, but that other factors (depending on the field) such as empathy, logical understanding, team skills, etc. play an essential role for a successful career as well. Accordingly, it is important to emphasise that grades do not determine an absolute career path.

References

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Genesee, F. (2006). *Educating English language learners: A synthesis of research evidence*. Cambridge University Press.

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4–27. https://doi.org/10.1037/0033-295X.102.1.4

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology, 74*(6), 1464–1480. https://doi.org/10.1037/0022-3514.74.6.1464

Hofmann, W., Gawronski, B., Gschwendner, T., Le, H., & Schmitt, M. (2005). A meta-analysis on the correlation between the implicit association test and explicit self-report measures. *Personality & social psychology bulletin*, *31*(10), 1369–1385. https://doi.org/10.1177/0146167205275613

International Labor Office. (2013). *International Standard Classification of Occupations 2008 (ISCO–08): structure, group definitions and correspondence tables*. International Labor Office.

Kessels, U. & Hannover, B. (2002). Die Auswirkungen von Stereotypen über Schulfächer auf die Berufswahlabsichten Jugendlicher. In Spinath, B. (Hrsg.) & Heise, E. (Hrsg.), *Pädagogische Psychologie unter gewandelten gesellschaftlichen Bedingungen: Dokumentation des 5. Dortmunder Symposions für Pädagogische Psychologie* (S. 53-67). Kovač.

Liu, W. C., Wang, C. K. J., & Parkins, E. J. (2005). A longitudinal study of students' academic self-concept in a streamed setting: The Singapore context. *British Journal of Educational Psychology*, *75*(4), 567–586. https://doi.org/10.1348/000709905X42239

Marsh, H. W. (1990). The structure of academic self-concept: The Marsh/Shavelson model. *Journal of Educational Psychology, 82*(4), 623–636. https://doi.org/10.1037/0022-0663.82.4.623

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic Self-Concept, Interest, Grades, and Standardized Test Scores: Reciprocal Effects Models of Causal Ordering. Child Development, 76(2), 397–416. https://doi.org/10.1111/j.1467-8624.2005.00853.x

Marsh, H. W., Abduljabbar, A. S., Parker, P. D., Morin, A. J. S., Abdelfattah, F., Nagengast, B., Möller, J., & Abu-Hilal, M. M. (2015). The Internal/External Frame of Reference Model of Self-Concept and Achievement Relations: Age-Cohort and Cross-Cultural Differences. *American Educational Research Journal*, *52*(1),168202. https://doi.org/10.3102/0002831214549453

Möller, J., Zitzmann, S., Helm, F., Machts, N., & Wolff, F. (2020). A meta-analysis of relations between achievement and self-concept. *Review of Educational Research, 90*(3), 376–419.
https://doi.org/10.3102/0034654320919354

Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2006). Self-esteem, academic self-concept, and achievement: How the learning environment moderates the dynamics of self-concept. *Journal of Personality and Social Psychology*, *90*(2), 334–349. https://doi.org/10.1037/0022-3514.90.2.334

VandenBos, G. R., & American Psychological Association (Hrsg.). (2015). *APA dictionary of psychology* (Second Edition). American Psychological Association.

Meta Analysis of Social Desirability Across Survey Modes

Dea Dautaj, Joyce Haler, Laura Heffenträger, Audrey Kontshakovski, Lea Müller, Hannah Streubert

Supervisors: Dr. Andreia Costa, Dr. Philipp Sischka, Christina Reisinger, Miriam Zimmer

Studies investigating the effect of different assessment modes on socially desirable response behaviour, come to dissimilar conclusions. Following on from this, the presented study was carried out to investigate the aforementioned topic in relation to the paper-pencil and online assessment modes. For this purpose, data from a total of eight studies with an overall sum of 4015 was analysed. Mostly college students were recruited, who had to be at least 18 years old and should not be diagnosed with any illness. The response behaviour was measured using the Balanced Inventory of Desirable Responding (BIDR) and the Marlowe-Crowne Social-Desirability Scale (MCSD) scales. In addition to the main effect investigated between assessment mode and socially desirable response behaviour, gender and publication year were also considered as potential factors correlating with social desirability. No significant effects were found neither for the main effect nor for the examined moderators.

1. Introduction

In the 21st century, computers and the Internet are more present than ever and are becoming increasingly important (Couper, 2011). Hardly any area of life is not affected by digitization. This also applies to a wide variety of tests and data collection in psychology. No matter how carefully one tries to proceed, there are certain variables that limit the validity of the results (Krebs, 1991). These also include social desirability. By definition, this contains, “on the one hand, a personality trait manifested in the need for social recognition, on the other hand, a situation-specific reaction to data collection, whereby actual facts are concealed or glossed over because of certain fears of consequences" (Stangl, 2021).

From the point of view of the "rational choice theory" (RCT), this behaviour can be explained by a conscious decision-making process in which individual costs and benefits of the behaviour are weighed up (Krebs, 1991). The theory assumes that an individually different need for recognition is an intrinsic motivation for behaviour and therefore has a decisive influence on the factor of social desirability. In addition, there is, as a cognitive basis for explanation, the expectation of external consequences. The extent of this depends on the way the study is conducted and the way data are collected. Depending on how anonymous and private the participating person assesses the treatment of their given data, they will answer in a socially desirable way in order to satisfy the need for recognition or to avoid negative reactions (Stocké, 2004). Buchanan (2000) already observed that social desirability is indispensably connected to privacy and anonymity. Related to RCT theory are the assumptions

about impression management (IM) (Paulhus, 1984) as also mentioned in the definition. This refers to various conscious techniques that are used to present oneself positively to others *(Impression Management,* n. d.). These are strongly context dependent. In addition, Paulhus (1984) speaks of self-deceptive enhancement (SDE), which can be regarded as a stable personality variable. Here, the focus is on striving to create or maintain a positive self-image of oneself.

Many studies have looked at assessment mode differences over the past decades, including the difference between online and paper-pencil. The results do not show a consistent trend and it is still not clear whether and if so, how different assessment modes affect the tendency to respond in a socially desirable way. Some studies (Joinson, 1999; Martin & Nagao, 1989) have concluded that the tendency to respond socially desirable decreases when surveys are taken in computer format. Buchanan's (2000) "Candor Hypothesis," which assumes the above-mentioned effect, supports this tendency. Contrary to these findings, Lautenschlager & Flaherty (1990) and Rosenfeld et al. (1996), for example, observed that the tendency for socially desirable responding increases in computer formats as opposed to paper-pencil. In turn, other studies failed to find any effects of administration modes (Booth-Kewley et al., 1992; Chua et al., 2006). A meta-analysis by Dodou & de Winter (2014) confirmed this assumption. Here, no effect of the different modes was found. However, several meta-analyses did not reach a coherent conclusion either. Weisband and Kiesler (1996) and Dwight and Feigelson (2000) concluded that socially desirable response tendencies decrease with computer surveys, in line with the Candor Hypothesis. However, Dwight and Feigelson (2000) distinguished between the two components of social desirability IM and SDE (Paulhus, 1984), finding a significant effect only for IM. These results were confirmed by Gnambs (2014). He found in his meta-analysis that individuals in computer-based surveys were 1.5 times more likely to answer questions about sensitive behaviour truthfully than those in PP. It should be noted, however, that these results were not measured by social desirability (SoD) scales but by willingness to talk about sensitive topics. Richman et al. (1999), did not find any effects of the modes per se, but only possible influences of the setting. They found that in an individual setting, less SoD occurred in computer than in paper- pencil (PP) for example. Many of the studies looking at mode differences also noted the perceived anonymity of the situation as a determining variable. Some studies found that there were significantly fewer socially desirable responses with increased anonymity (Joinson, 1999; Rosenfeld et al., 1996; Gnambs & Kaspar, 2014). However, others found no significant effect (Dodou & de Winter, 2014; Richman et al, 1999; Fox & Schwartz, 2002). As seen, the results are contradicting. Therefore, we will conduct a meta-analysis concerning possible mode effects.

Additionally, gender may also be a potential factor for differences in the tendency to respond in a socially desirable manner. Gnambs and Kaspar (2014) and Pope-Davis and Twing (1991) found no significant differences of gender on socially desirable responding in different modes. Krebs (1991) also noted that the tendency to respond socially desirable was the same for males and females. However, there are many studies that indicate attitudinal differences between men and women when it comes to computers. Women are more sceptical of computers, spend less time with them, and arguably take longer overall to learn about and become comfortable

with such new devices than men (Miles & King Jr., 1998; Broos, 2005). Women are also usually found to have higher scores for internet anxiety (Broos, 2005; Meier & Lambert, 1991). It is interesting to note that both a study by Joiner et al. (2005) and a study by Miles and King (1998) found no gender differences in computer anxiety and computer knowledge among college students. This could be explained by the fact, that students nowadays, regardless of gender, have grown up with and are therefore accustomed to computers and use them in university (Miles & King, 1998).

Also, it is hypothesized that mode effects on social desirability responses change according to the publication year. A large number of studies have observed a decline in the "Candor Hypothesis" (Buchanan, 2000) over the years (Dodou & de Winter, 2014; Gnambs & Kaspar, 2014; Booth-Kewley et al., 1996). Dwight and Feigelson (2000) note that in the 1990s, effect sizes were almost always around zero and suggest, for example, as a potential reason that understanding of a computer's capabilities has improved over the years. This goes along with the so-called "Big Brother Syndrome" (Martin & Nagao, 1989), in which the anonymising effect of computers is assumed to be reversed by a higher understanding of various "surveillance capabilities" of a computer and the Internet. A study by Rosenfeld et al. (1996) talks about the fact that nowadays the message people get from computers has more and more to do with surveillance and monitoring. This, of course, has implications for the perceived anonymity and privacy already mentioned. Since people nowadays are much more familiar with the fact that computer networks are interconnected and their answers can be quickly checked, they pay more attention to truthful answers when there is a possibility for verification, according to Martin & Nagao (1989). If this is not the case, as for example in SoD scales, which have no "true" values and measure a latent construct (Tourangeau & Yan, 2007), socially desirable responding increases again. However, Gnambs & Kaspar also noted in their 2014 meta-analysis that effects that support less socially desirable responding in computer formats have increased again in the last decade, which may be due in part to habituation effects and greater acceptance of computer-based psychological testing procedures. Due to the given empirical evidence, no clear direction can be identified so far, which is why we will investigate this phenomenon again in the present meta-analysis. In contrast to many other meta-analyses, we will focus exclusively on persons aged 18 or older. This decision is based on the fact that the questionnaires used, meaning the BIDR and the MCSD scale, are validated only for persons aged 18 and older. In a study from 1991, Krebs found that some age ranges are more prone to SoD answering than others. Another study by Mwamwenda (1995) found that adolescents show significantly higher SoD tendencies than adults. It is therefore reasonable to assume that the inclusion of participants younger than 18 would alter the results. This is an influencing factor that can be eliminated by adult-only participants.

Furthermore, in this meta-analysis, only studies that use already validated scales such as the BIDR or the MCSD are being used. Dwight and Feigelson (2000) for example, noted that the many different results could be due to the fact that no uniform instruments were used to measure social desirability.

Accordingly, with our meta-analysis, we hypothesize that (1) assessment modes have an effect on the tendency to answer socially desirable. Furthermore, because various studies talk about sex differences in

attitudes towards computers, a potential effect of gender will be investigated. We assume (2) that women tend to answer more socially desirable in computer surveys than men. Because of possible changes regarding the opinion about computers and the internet (Big Brother Syndrome) we further hypothesize that (3) the year of publication has an effect on the use of different assessment modes and socially desirable responding.

2. Methods

2.1 Procedure

The literature was conducted via multiple databases, namely PsycInfo and GoogleScholar. This was done via a search string using keywords such as *social desirability*, *sensitive questions* in combination with *paper-pencil*, *web-based*, *computerized questionnaire* and other synonyms in relation to this topic (see Appendix A1).

Once the database was complete, all studies were transferred into CADIMA. CADIMA is a free web tool facilitating the conduct and assuring for the documentation of systematic reviews, systematic maps, and further literature reviews.

It is to note that this project is part of a larger network-meta-analysis, where all possible administration modes are being investigated and compared. Therefore, it includes a fairly large number of papers. The comparison between PP and computer is only a small part of this larger project, which is why not all of the found studies relate to this comparison specifically.

A study was included in the meta-analysis when it fit the following criteria: (a) The fulltext is available in English language. (b) The publication type is a journal article. (c) If a score of social desirability for either a paper-pencil or a computer-based questionnaire is reported with one of the following social desirability scales: the Balanced Inventory of Desirable Responding, the Marlowe Crowne Social Desirability Scale, L- or K-scales of the Minnesota Multiphasic Personality Inventory. The short versions of each questionnaire were acceptable as well if they had been approved. (d) The study contains the following Statistical metrics: (d.1) the sample size for both compared groups, mean (M) for the estimated social desirability on the respective questionnaire for or standard error (SE) of both compared groups or full sample standard deviation; (d.2) Hedges' g for the estimated social desirability comparison between both groups; (d.3) Cohens d for the estimated social desirability comparison between both groups, total sample size; (d.4) *T*-value of t-test or its *p*-value for the estimated social desirability comparison between both groups, sample size for both compared groups or total sample size; (d.5.) F-value (one-way-ANOVA) for the estimated social desirability comparison between both groups, sample size for both compared groups or total sample size; (d.6.) Unstandardized or standardized regression coefficient for the estimated social desirability comparison between both groups, sample size for both compared groups, SD of group comparison.; (d.7.) Either the ANCOVA F-value for the estimated social desirability comparison between both groups and sample size for both compared groups and (covariate outcome correlation or multiple correlation) and number of covariates, the adjusted mean from ANCOVA for both compared groups and adjusted or pooled standard deviation and sample size for both compared groups and (covariate outcome correlation or multiple correlation) and number of covariates, the ANCOVA p-value (one or two-tailed) and

sample size for both compared groups and (covariate outcome correlation or multiple correlation) and number of covariates and tailedness of ANCOVA p-value (one or two tailed) or the ANCOVA t-test and sample size for both compared groups and (covariate outcome correlation or multiple correlation) and number of covariates; (d.8.) Chi-Square coefficient for the estimated social desirability comparison between both groups or the p-value of the chi-square or phi-square and a vector of total sample size(s). (e) An experimental design was employed. (f) The study includes only participants with minimum age of 18. (g) The study includes a comparison of the assessment modes paper-pencil and online that reports a score for social desirability on the mentioned questionnaires.

The study was excluded if one of the following criteria was fulfilled. (h) If study participants needed to be diagnosed with a preexisting physiological or psychological illness. (i) The participants were exclusively recruited from a forensic population. (j) The study explicitly stated interventions to prevent social desirable responding, not meaning skipping of elements or backtracking in the questionnaire nor the choice of answering setting or proctoring. (k) There is no comparison after controlling for the comparison exclusion criterion for instance after using an intervention for preventing social desirability.

Furthermore, all studies were screened by two different group members first by title and abstract and if not excluded by then, by fulltext to find eligible studies.

For the data analysis, the Cohen's d effect size was used. It is a standardized effect size for mean differences (Cohen, 1988).

After the results showed that the heterogeneity was overly high, it was decided to use the random effects model. Contrary to the fixed effects model, it assumes that the average standard error in the population varies from study to study not only because of random error but also because of true variation in effect sizes from study to study (Borenstein et al., 2007).

$$Cohen's\ d = \frac{M1-M2}{SDpooled} \quad \text{, where} \qquad SDpooled = \sqrt{\frac{(SD1^2+SD2^2)}{2}}$$

2.2 Materials

Social desirability bias can be measured through means of formal scales or through the responses to follow-up questionnaires. These methods contain sensitive questions that are designed to examine if the participants answer socially desirable (Dodou & de Winter, 2014). These contain sensitive questions to see if the participants have a response bias concerning social desirability. Richman et al. (1999) recognized that *instruments specifically designed to measure social desirability distortion […] are more reliable and direct indicators of distortion than instruments used to measure diverse other traits, syndromes, attitudes, and behaviour* (p.757). This supports the choice to only use validated scales in the meta-analysis.

The MCSD (Crowne and Marlowe, 1960) and the BIDR (Paulhus, 1984) scales were used as measurement instruments. Although not given in the selected studies, the L and K scales of the Minnesota Multiphasic Personality Inventory (MMPI), could have been used.

The MCSD scale is a self-assessment questionnaire and includes 33 true or false items. In 1960, it was developed by

Crowne and Marlowe to measure self-assessment bias due to social desirability, as this is considered one of the most common difficulties in surveys. An illustration of this scale is: “I am always careful about my manner of dress” or “I’m always willing to admit it when I make a mistake.” which are the items 7 and 16.

The BIDR was developed in the same way as the MCSD as Measurement and control of response bias. It was developed by Paulhus (1984). It comprises a total of 40 items and 2 subscales. The first subscale is the Self-Deceptive Enhancement subscale, which assesses whether the respondent's answer is honest and sincere. An example of this scale is: "I am a completely rational person". The second subscale is Impression Management, i.e., the tendency for social recognition. "Once in a while I laugh at a dirty joke" is a typical example for this scale.

3. Analysis

Besides the moderators publication year and gender proportion, setting was chosen as a potential confounding variable, as well.

For instance, in a laboratory setting, there should be no confounding variables present, whereas these can appear in a non-lab setting, e.g., at the university in the presence of other students. To analyse the effect of this moderator, we coded laboratory with 1 and non-laboratory with 0.

Gender proportion was chosen as another moderator variable to analyse if the majority of the participants being women affects our data and the effect size in general. The proportion was coded as follows: more women in the studies were coded as 1, and more men as 0.

To analyse if significant effects can be found, the effect size of Cohen’s d was used. The effect sizes can be understood as follows: a positive effect would suggest higher SoD scores in the computer assessment mode, whereas a negative effect would mean that the SoD scores are higher in the paper-pencil mode.

To perform the analysis, we used the Comprehensive Meta-Analysis Software (CMA). Beginning with the transfer of our data from the selected studies into the program, we then ran the analysis to calculate our main effect size and examined the heterogeneity in the observed effect size. Moreover, a multiple regression analysis with the three moderator variables was conducted.

In addition, a potential publication bias was investigated. If this were the case, it would affect the precision and correctness of the calculated effects. To examine the potential presence of a publication bias, we used the trim and fill method (Duval & Tweedie, 2000).

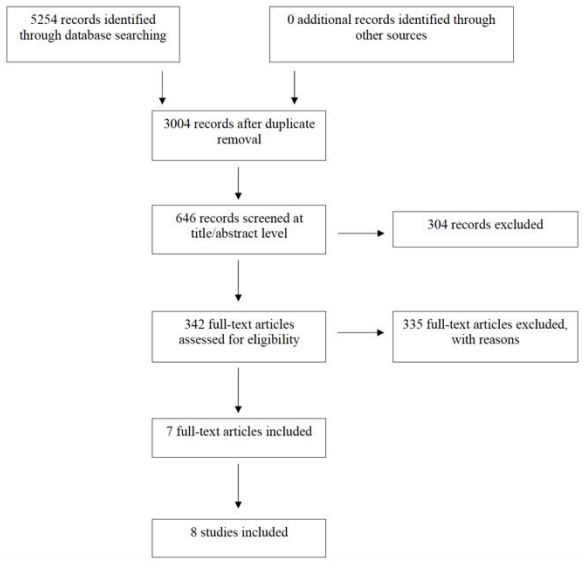
4. Results

Fig. 1 reveals the flow diagram of the study selection. The use of the PsycINFO and Google Scholar search operators resulted in 5254 identified papers. 3004 remained after removing all the duplicates. 646 were screened by their titles and abstracts. 304 articles were excluded after the screening, not fulfilling the inclusion criteria. The remaining 342 records were then reviewed by reading their full-text articles. 335 studies were finally excluded, not fitting our criteria. All in all, 7 studies’ full-text articles were included. One full-text article reported results from two different studies (Study 1 + Study 2), which were treated independently. Consequently, we conducted

the meta-analysis with 8 different studies in total.

Figure 1

Flow diagram depicting the study selection process



4.1 Main effect size

The main outcome is an almost null adverse effect (d = - .019, p = .939), which is not significant (*Fig*. 2). After examining the heterogeneity, we received an overly high and significant value for Cochran's Q: Q = 237.738 for the fixed effects model.

Figure 2

Forest Plot displaying results

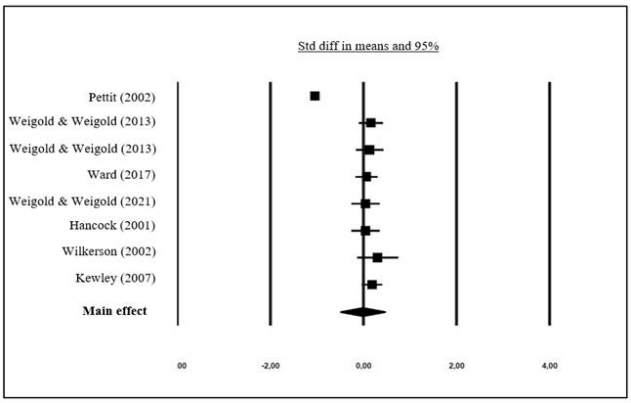


Table 1

Meta-Analysis results

Study name	Std. means	Std. error	Variance	Lower limit	Upper limit	z- value	p-value
A comparison of World-Wide Web and paper-and-pencil personality questionnaires	-1,040	0,055	0,003	-1,147	-0,933	-19,073	0,000
Examination of the equivalence of self-report survey-based paper-and-pencil and internet data collection methods - Study 1	0,163	0,126	0,016	-0,085	0,410	1,290	0,197
Examination of the equivalence of self-report survey-based paper-and-pencil and internet data collection methods - Study 2	0,134	0,143	0,020	-0,147	0,414	0,934	0,350
Paper/pencil versus online data collection: An exploratory study	0,068	0,123	0,015	-0,172	0,309	0,558	0,577
Computerized Device Equivalence: A Comparison of Surveys Completed Using A Smartphone, Tablet, Desktop Computer, and Paper-and-Pencil	0,043	0,159	0,025	-0,268	0,354	0,273	0,785
Comparing Social Desirability Responding on World Wide Web and Paper-Administered Surveys	0,044	0,150	0,023	-0,250	0,338	0,294	0,769
Socially Desirable Responding in Computerized Questionnaires: When Questionnaire Purpose Matters More Than the Mode	0,303	0,218	0,048	-0,124	0,731	1,390	0,165
Social desirability effects on computerized and paper-and-pencil questionnaires	0,186	0,116	0,013	-0,040	0,413	1,610	0,107
Overall	-0,019	0,245	0,060	-0,498	0,461	-0,076	0,939

4.2 Regression analysis with moderator variables

The standardized difference in means for the laboratory setting was not significant, with a value of d = .133, p = .299. The same applies for the standardized difference in means for the non-laboratory setting being d = - .127, p = .699. Not all studies reported exclusively one setting. Therefore, one study could not be coded properly. This has to be taken into account when looking at the overall result. The overall effect of the setting is small and not significant with d = .129, p = .137.

The regression analysis with the moderator *publication year* showed a positive, but almost null b-value of the regression equation, which means the younger the study, the higher the standardized difference in means. However, this effect was non-significant (p > .05).

Furthermore, regarding the moderator *gender proportion*, less women among the participants lead to a higher standardized difference in means (b = - .561).

4.3 Publication bias

Table 2 shows that the trim and fill method resulted in a remarkable difference between the observed and adjusted values, which means there is a publication bias present in this meta-analysis.

Table 2

Duval and Tweedie's trim and fill for random effects model

	Studies trimmed	Point Estimate	Lower Limit	Upper Limit
Observed values		- 0.01863	- 0.49848	0,46122
Adjusted values	5	- 0.70745	- 0.17714	-0,23776

5. Discussion

The goal of the presented meta-analysis was to examine differences of social desirability across paper-pencil and computer questionnaires, based on 8 studies published over the last 20 years. In addition, analyses on the moderator variables gender, publication year and setting were conducted.

This meta-analysis is limited in scope, since it is a small part of a larger network-meta-analysis and was done as part of an experimental internship, which led to a limited time. Additionally, the strict inclusion and exclusion criteria do exclude a considerable part of studies, but bring the advantage that ambiguities and potential influencing factors, such as unvalidated scales, under 18-year-olds or people with diagnosed mental/physical illness, can be circumvented. This allows for better comparability and control of certain factors and valuable insights were gained that can be included in a larger scale in future research.

Our first hypothesis aimed to investigate whether assessment modes, in this case, PP and computer, have an effect on the tendency to answer socially desirable. No significant effects were found and thus the "Candor Hypothesis" (Buchanan, 2000) could not be confirmed. The results of this analysis reinforce the findings of a meta-analysis by Dodou and de Winter (2014) who found no effect of administration modes. This could be explained by the fact that the two administration modes have become increasingly alike, as well as participants being accustomed to both formats (Dodou & de Winter, 2014; Tourangeau et al., 2007). A computer in the early days was new and exciting, a neutral device without any social judgement (Dodou & de Winter, 2014). This novelty is long gone and attitudes towards computers have also changed e. g. due to the "Big Brother Syndrome" (Martin & Nagao, 1989). Perhaps, the initial "hype" about computer formats has simply died down over the last decades. Hence, it is possible that the two modes do not show many differences anymore as both are normality and established in the world of testing.

Very small and non-significant results were found for the moderator analyses. The second hypothesis, that women tend to answer more socially desirable than men, could not be confirmed, since there were no significant effects for the moderator *gender proportion*. These results reflect the findings of Krebs (1991) who found no effect for gender either. As mentioned in the Introduction, no gender differences could be found in younger people, specifically students, with regard to attitudes towards computers (Joiner et al., 2005; Miles & King, 1998). Since the absolute majority of participants in this meta-analysis were also young students, this may explain why no significant differences were found here. Since younger people in particular, regardless of gender, are more accustomed to computers and little to no gender differences in perceived computer skills were found (Varank, 2007), it is acceptable that

potential variations, if any, could be found in older individuals. It is recommended to look at an interaction between gender and age and analyse if this would produce a different outcome.

The third hypothesis, that the year of publication has an influence on the different findings, could not be confirmed either. Unlike some other meta-analyses and studies (Dodou & de Winter, 2014; Dwight & Feigelson, 2000; Gnambs & Kaspar, 2014), no significant effect of this moderator was found. This may be due, in part to the fact that some of these analyses used a much broader publication year span, including studies from the 1980s and 1990s (Gnambs & Kaspar, 2014; Dwight & Feigelson, 2000), where computers and their use were not as common and rather unknown territory for many people. This is particularly interesting given that the studies used for this meta-analysis consisted only of studies published in the years 2001 to 2021. Arguably, the time period used in this meta-analysis is more representative for the current situation. There is potentially no significant difference between the two modes, since computers are no longer new and are part of everyday life. Older studies could lead to a distortion of the actual effect, as attitudes and habituation to the different formats naturally change over decades (Gnambs & Kaspar, 2014) and thus do not reflect the present situation.

5.1 Limitations and outlook

This meta-analysis has a number of limitations that have to be taken into account when looking at its practical implications.

Due to the limited nature of this project, only a handful of studies could be extracted. With merely eight papers, this analysis cannot be taken as representative for the entirety of this topic. It is not unlikely that a larger number of studies would have yielded a different result. Due to this low number of studies, heterogeneity is too high, even with the random effects model. More studies would most likely lead to more homogeneity and thus to a better comparability of the different study results. Additionally, a publication bias could be reduced or removed with more added studies.

Furthermore, research has proven higher levels of socially desirable responding to be associated with advancing age (Soubelet & Salthouse, 2011). Of the 8 studies included in our meta-analysis, 7 exclusively consisted of university students. As the samples mainly comprised of young participants, age differences in socially desirable responding across survey modes were not taken into consideration. It is highly recomended to examine a wider age range in future research.

Besides, effect sizes were measured with Cohen's d, which contrary to Hedge's g, does not account for different sample sizes. In the context of this meta-analysis Cohen's d was considered to be the optimal choice to avoid further study exclusions since a clear number of participants was not indicated in two studies and the authors could not be reached. However, in a larger setting and with more time, Hedge's g would be recommended to avoid loss of information and possible bias in the results.

Systematic cross-cultural differences in SoD have been shown (Johnson & Vijver, 2003), however, not addressed in this specific work. Cultural discrepancies in socially desirable responding across different survey modes are recommended to analyse in future research.

Moreover, an extension of moderator variables as for example anonymity can

be suggested to future researchers because they might have a decisive influence on social desirability tendencies in different modes (Dodou & de Winter, 2014).

Practical implications:

The question of whether administration modes have a significant effect on response tendencies is of particular interest to researchers who need to decide which mode to use. In accordance with the results from Dodou and de Winter's meta-analysis (2014), it can be assumed that computer-based questionnaires, through financial and logistical advantages, offer a higher ecological validity, since there is virtually no difference for the tendency to answer socially desirable. It is important however, to remember that this is a small meta-analysis and future research is needed.

5.2 Conclusion

This meta-analysis has pointed out that all effect sizes in the moderator analyses were small, and not statistically significant. It can therefore be concluded that in regards to socially desirable responding the choice of assessment modes, whether computer or paper- pencil, does not result in any systematic difference. The same conclusion applies for the moderator variable publication year. However, in regards to gender we were able to exclude an effect of women answering more socially desirable than men. More research on this topic needs to be done since our results are limited to a small number of studies.

6. References

Booth-Kewley, S., Edwards, J. E. & Rosenfeld, P. (1992). Impression management, social desirability, and computer administration of attitude questionnaires: Does the computer make a difference? *Journal of Applied Psychology*, *77*(4), 562–566. https://doi.org/10.1037/0021-9010.77.4.562

Broos, A. (2005). Gender and Information and Communication Technologies (ICT) Anxiety: Male Self-Assurance and Female Hesitation. *CyberPsychology & Behavior*, *8*(1), 21–31. https://doi.org/10.1089/cpb.2005.8.21

Borenstein, M., Hedges, L. V. & Rothstein, H. R. (2007). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. https://doi.org/10.1002/jrsm.12

Buchanan, T. (2000). Potential of the Internet for Personality Research. *Psychological Experiments on the Internet*, 121–140. https://doi.org/10.1016/b978-012099980-4/50006-x

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Chuah, S. C., Drasgow, F. & Roberts, B. W. (2006). Personality assessment: Does the medium matter? No. *Journal of Research in Personality*, *40*(4), 359–376. https://doi.org/10.1016/j.jrp.2005.01.006

Couper, M. P. (2011). The Future of Modes of Data Collection. *Public Opinion Quarterly*, *75*(5), 889–908. https://doi.org/10.1093/poq/nfr046

Crowne, D. P. & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*(4), 349–354. https://doi.org/10.1037/h0047358

Dodou, D. & de Winter, J. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior*, *36*, 487–495. https://doi.org/10.1016/j.chb.2014.04.005

Duval, S. & Tweedie, R. (2000). Trim and Fill: A Simple Funnel-Plot-Based Method of Testing and Adjusting for Publication Bias in Meta-Analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341x.2000.00455.x

Dwight, S. A. & Feigelson, M. E. (2000). A Quantitative Review of the Effect of Computerized Testing on the Measurement of Social Desirability. *Educational and Psychological Measurement*, *60*(3), 340–360. https://doi.org/10.1177/00131640021970583

Fox, S. & Schwartz, D. (2002). Social desirability and controllability in computerized and paper-and-pencil personality questionnaires. *Computers in Human Behavior*, *18*(4), 389–410. https://doi.org/10.1016/s0747-5632(01)00057-7

Gnambs, T. & Kaspar, K. (2014). Disclosure of sensitive behaviors across self-administered survey modes: a meta-analysis. *Behavior Research Methods*, *47*(4), 1237–1259. https://doi.org/10.3758/s13428-014-0533-4

Gnambs, T. & Kaspar, K. (2016). Socially Desirable Responding in Web-Based Questionnaires: A Meta-Analytic Review of the Candor Hypothesis. *Assessment*, *24*(6), 746–762. https://doi.org/10.1177/1073191115624547

Heterogeneity in Meta-analysis (Q, I-square) - StatsDirect. (o. D.). Www.Statsdirect.Com.

https://www.statsdirect.com/help/meta_analysis/heterogeneity.htm

Johnson, T. P., & Van de Vijver, F. J. (2003). Social desirability in cross-cultural research. *Cross-cultural survey methods*, *325*, 195-204.

Joiner, R., Gavin, J., Duffield, J., Brosnan, M., Crook, C., Durndell, A., Maras, P., Miller, J., Scott, A. J. & Lovatt, P. (2005). Gender, Internet Identification, and Internet Anxiety: Correlates of Internet Use. *CyberPsychology & Behavior*, *8*(4), 371–378. https://doi.org/10.1089/cpb.2005.8.371

Joinson, A. (1999). Social desirability, anonymity, and internet-based questionnaires. *Behavior Research Methods, Instruments & Computers*, *31*(3), 433–438. https://doi.org/10.3758/bf03200723

Julius Kühn-Institut, Federal Research Centre for Cultivated Plants. (2012). *CADIMA*. CADIMA. https://www.cadima.info/index.php

Krebs, D. (1991). *Was ist sozial erwünscht? Der Grad sozialer Erwünschtheit von Einstellungsitems.* (ZUMAArbeitsbericht,1991/18).

Mannheim: Zentrum für Umfragen, Methoden und Analysen -ZUMA

https://nbnresolving.org/urn:nbn:de:0168-ssoar-69010

Lautenschlager, G. J. & Flaherty, V. L. (1990). Computer administration of questions: More desirable or more social desirability? *Journal of Applied Psychology*, *75*(3), 310–314. https://doi.org/10.1037/0021-9010.75.3.310

Lexikon der Psychologie- Impression Management. (o. D.). spektrum. Abgerufen am 30. November 2021, von https://www.spektrum.de/lexikon/psychologie/impression- management/7066

Martin, C. L. & Nagao, D. H. (1989). Some effects of computerized interviewing on job applicant responses. *Journal of Applied Psychology*, *74*(1), 72–80. https://doi.org/10.1037/0021-9010.74.1.72

Mcleod, S. (2019, 10. Juli). *Effect Size*. Www.Simplypsychology.Org. https://www.simplypsychology.org/effect-size.html

Meier, S. T. & Lambert, M. E. (1991). Psychometric properties and correlates of three computer aversion scales. *Behavior Research Methods, Instruments & Computers*, *23*(1), 9–15. https://doi.org/10.3758/bf03203329

Miles, E. W. & King, W. C. (1998). Gender and Administration Mode Effects when Pencil-And-Paper Personality Tests are Computerized. *Educational and Psychological Measurement*, *58*(1), 68–76. https://doi.org/10.1177/0013164498058001006

Mwamwenda, T. S. (1995). Age Differences in Social Desirability. *Psychological Reports*, *76*(3), 825–826. https://doi.org/10.2466/pr0.1995.76.3.825

Paulhus, D. L. (1984). Two-component models of socially desirable responding. *Journal of Personality and Social Psychology*, *46*(3), 598–609. https://doi.org/10.1037/0022-3514.46.3.598

Pope-Davis, D. B. & Twing, J. S. (1991). The effects of age, gender, and experience on measures of attitude regarding computers. *Computers in Human Behavior*, *7*(4), 333–339. https://doi.org/10.1016/0747-5632(91)90020-2

Richman, W. L., Kiesler, S., Weisband, S. & Drasgow, F. (1999). A meta-analytic study of social desirability distortion in computer-administered questionnaires, traditional questionnaires, and interviews. *Journal of Applied Psychology*, *84*(5), 754–775. https://doi.org/10.1037/0021-9010.84.5.754

Rosenfeld, P., Booth-Kewley, S., Edwards, J. E. & Thomas, M. D. (1996). Responses on computer surveys: Impression management, social desirability, and the big brother syndrome. *Computers in Human Behavior*, *12*(2), 263–274. https://doi.org/10.1016/0747-5632(96)00006-4

Soubelet, A., & Salthouse, T. A. (2011). Influence of social desirability on age differences in self-reports of mood

and personality. *Journal of personality*, *79*(4), 741-762.

Stangl, W. (2021). *Soziale Erwünschtheit*. Online Lexikon für Psychologie und Pädagogik. https://lexikon.stangl.eu/1807/soziale-erwuenschtheit

Stocké, V. (2004). Entstehungsbedingungen von Antwortverzerrungen durch soziale Erwünschtheit Ein Vergleich der Prognosen der Rational-Choice Theorie und des Modells der Frame-Selektion*, Zeitschrift für Soziologie*, Heft 4, 303- 320. https://doi.org/10.1515/zfsoz-2004-0403

Tourangeau, R. & Yan, T. (2007). Sensitive questions in surveys. *Psychological Bulletin*, *133*(5), 859–883. https://doi.org/10.1037/0033-2909.133.5.859

Varank, I. (2007). Effectiveness of Quantitative Skills, Qualitative Skills, and Gender in Determining Computer Skills and Attitudes: A Causal Analysis. *The Clearing House*, *81*(2), 71–80. https://doi.org/10.3200/TCHS.81.2.71-80

Weisband, S. & Kiesler, S. (1996). Self disclosure on computer forms. *Proceedings of the SIGCHI conference on Human factors in computing systems common ground - CHI '96*. https://doi.org/10.1145/238386.238387

7. Annexe

Appendix A: Search String

"advanced search, limit to: (English language and abstracts and "0100 journal") ((((interview or interrogation).af. or ((face to face)).af.) and (Telephone or phone).af.) or (((interview or interrogation).af. or (face to face).af.) and (Paper* or letter or post or mail or print or written or writing).af.) or (((interview or interrogation).af. or (face to face).af.) and (web or online or offline or internet or e-mail or computer*or electronic or mobile* or smartphone or iPhone or cellphone or android or Mac or OS or tablet or iPad or note* or laptop or PC).af.) or ((Telephone or phone) and (Paper* or letter or post or mail or print or written or writing)).af. or ((Telephone or phone) and (web or Online or offline or internet or e-mail or computer* or electronic or mobile* or smartphone or iPhone or cellphone or android or Mac or OS or tablet or iPad or note* or laptop or PC)).af. or ((Paper* or letter or post or mail or print or written or writing) and (web or online or offline or internet or e-mail or computer* or electronic or mobile* or smartphone or iPhone or cellphone or android or Mac or OS or tablet or iPad or note* or laptop or PC)).af.) and (((social* adj desirab*) or (social adj decontextuali*)).af. or disclosure.af. or (response adj choice).af. or (response adj pattern).af. or (impression adj management).af. or (response adj distortion).af. or (self adj deceptive adj enhancement).af. or Aggravation.af. or Dissimulation.af. or Faking.af. or Malinger*.af. or Simulation.af.) and ((Balanced adj Inventory adj of adj Desirable adj Responding).cv. or BIDR.cv. or (Paulhus adj deception adj scale).cv. or (Marlowe adj Crowne).cv. or MCSDS.cv. or (Minnesota adj Multiphasic adj Personality adj Inventory).cv. or MMPI.cv.)"

The connection between language, emotion and cognitive performance

Laure Wagner, Marie Sjöström, Lea Büth, Zoe Schneider, Fenja Degener, Zoé von Kraewel

Dr. Andreia Costa, Maïte Franco

The central question of this study was whether cognitive performance, more specifically, working memory performance differs depending on whether emotion is expressed in the mother tongue or in the second language. The exploratory hypothesis is based on Lindquist et al.'s findings (2015b) that the language we speak influences our emotionality. Another key theory is the Dual Model of Language and Cognition by Perlovsky (2009a) from which it is derived that emotion regulation capabilities should be differently developed in the mother tongue. As emotion regulation influences cognitive performance (Richards and Gross, 2000), we investigated whether the language in which emotion is expressed influences cognitive performance. An emotion was elicited with an unsolvable puzzle and we tested working memory. The methods used were questionnaires regarding the overall state of emotion, emotion regulation methods, tasks eliciting emotion, an interview to express the emotion and working memory tests. No significant difference in working memory performance was found between the experimental and control group. From that can be concluded that the language in which we express our emotions and the emotional regulation associated with this expression did not affect cognitive performance in our group.

Introduction

Emotions are self-governed and subconscious bodily responses to external factors (Izard, 2010) and contemporary models locate their origin in brain systems responsible for evaluating significant stimuli concerning our goals and needs (Ochsner & Gross, 2008). They serve the purpose of addressing problems, of providing information and they prepare -among others- for a rapid response (Gross, 1998). Emotion regulation is a process which affects the emotions we have and when and ho we experience them (Gross, 1998). According to Gross (1998), this process can be conscious or subconscious, automatic or controlled, and involves a change in the dynamics of emotions and the way we respond to situations. Festinger and Carlsmith (1959) state that we regulate our emotions when faced with a situation that elicits cognitive dissonance, since this unpleasant state results in an inner drive to restore cognitive harmony and to avoid disharmony. Furthermore, individuals differ in their trait-emotion regulation as there are dozens of strategies which can be used to regulate different emotions (Richards & Gross, 2000). In his Process Model of Emotion regulation, Gross (1998) distinguishes between four antecedent-focused strategies and one response modulation. The antecedent-focused strategies can be applied during various points of the process of emotion generation. They consist of situation selection, situation modification, attentional deployment, and cognitive change (Gross, 1998). The fifth strategy of the model, however, the response modulation, is only applied when an emotion has already been evoked.

Emotions and cognition are tightly linked such that emotion regulation strategies may impact cognitive capacities (Gross, 1998). Richards and Gross (2000), for example, found that instructing participants to use suppression as an emotion regulation strategy during an emotion eliciting situation impaired their incidental memory. In their study, those participants who were told to suppress their emotions performed worse in tasks involving cued recall and cued recognition compared to those who were not given any instructions as to how to regulate their emotions. Similarly, excessive use of suppression was observed to be related to the impairment of verbally encoded memory (Richards & Gross, 2000). However, Richard and Gross (2000) could not observe the same detrimental effect of suppression on memory when non-verbal information had to be memorized and recalled. The impaired memorization of verbal information could be explained by the fact that the use of suppression as an emotion regulation strategy requires self-focus and self-monitoring. Both use upper cognitive resources, which are also required for encoding and remembering events. In contrast, reappraisal does not seem to significantly impair cognitive processes. Richards and Gross (2000) surmise that this could be due to the fact that re-evaluating the emotion itself and re- assessing its importance could use less cognitive resources. Finally, it could be assumed that the intensity of the emotions felt influences the degree of memory impairment. Nevertheless, both seem to be independent, according to Richards and Gross (1999).

According to the Dual Model of Language and Cognition by Perlovsky (2009a) all concepts in our minds consist of both language and cognition. These are two separate, but linked systems

in which the concrete linguistic part overshadows the vaguer cognitive part from our consciousness (Perlovsky, 2009a). Franklin et al. (2008) showed that whenever adults learn a new word for a color, their left-brain hemisphere, known to process language, is activated instead of their right, which is responsible for visual perception. In infants however, one would observe an activation in the right hemisphere instead. Lindquist et al. (2015b) suggest that categorizing sensations using language is the process resulting in emotions. Categorization develops using emotion category knowledge which is supported by language (Lindquist & al., 2015a). In the course of an individual's life, language models expand faster than cognitive concepts. Therefore, with more experience we learn to fill in the words we learned with actual meaning. For example, when a child learns the word "motivation" he or she does not know its concrete implications and contents but in the process of growing older, they develop this knowledge (Perlovsky, 2009a). Thus, according to Perlovsky, it is necessary to learn abstract language concepts first as an anchor for cognitive concepts. He then argues that the words we ground our cognitive models in determine our cognitive concepts (Perlovsky, 2009a). Consequently, which language we speak determines the shape of our cognitive models. This is often referred to as the Sapir-Whorf Hypothesis: the language we speak influences how we think (Whorf, 1956). Assuming that additional time to develop language skills leads to a higher competence level in cognition and therefore in emotion regulation as well, it can be concluded that a second or newly learned language would not display the same level of emotion regulation capability.

Perlovsky (2009b) further put the Sapir- Whorf hypothesis into the context of emotions. He stated that different languages have different levels of emotionality linked to the sound of the words. Even though language is often considered separate from primitive emotion, the connection between emotion and meaning through sound remains (Perlovsky, 2009b). For example, in a study by Guttfreund (1990) bilinguals showed a different emotional intensity depending on what language they were interviewed in. Participants interviewed in Spanish showed more intense emotions than participants interviewed in English. This effect did not depend on whether the language was their mother tongue or their second language (Guttfreund, 1990). This uniquely relates to the connection between the sound of the words and the language's emotionality and makes no statement concerning the link between language and cognition or emotion regulation (Guttfreund, 1990). This complements the findings by Lindquist et al. (2015b), that people across cultures perceive emotions differently depending on the language they speak.

Exploring the effects of emotion, Guttfreund refers to the Rozensky and Gomez study (1983) about language effects on bilingualism, using Spanish as first language and English as a second. The effect states that patients communicating in a second language seem emotionally more withdrawn compared to communication in their mother tongue. It was observed that an emotional barrier and an increase in defensive behavior took place, when bilinguals engaged in a second language (Rozensky & Gomez, 1983 as in Guttfreund, 1990). This effect is retraceable to the increased use of cognitive capacity, forcing an intellectualized approach to the emotional expression, and suppressing the affective component (Rozensky & Gomez, 1983). Therefore, expressing one's emotions in a second language instead of the mother tongue could lead to a different emotion regulation strategy. Since the participants seemed emotionally more withdrawn, as if erecting an emotional barrier, this could be either suppression or a healthy distancing from one's emotions. This effect was called emancipatory detachment by Kellman (2000).

The bilingual advantage hypothesis states that bilinguals better select and control the focus of their attention because they are used to switching between languages (Bialystok, 2011, 2017; Bialystok & al., 2012; Kroll & Bialystok, 2013). With practice, this is supposed to become a general effect in their cognitive patterns (Lowe & al., 2021). However, according to Lowe et al.'s (2021) meta-analysis, even though the effect of bilingualism on children's executive functioning was statistically significant, it was small and influenced by extraneous variables, such as socioeconomic status. Moreover, a bilingual advantage only appeared in studies with verbal tasks (Lowe & al., 2021). Therefore, the mere fact that someone can speak several languages should barely influence cognitive performance and have no influence at all on the performance in non-verbal cognitive tasks.

In summary, the Dual Model of Language and Cognition by Perlovsky (2009a) and the Sapir-Whorf hypothesis (Whorf, 1956) state that the language we speak influences how we think. Furthermore, as can be derived from the Emotional Sapir-Whorf hypothesis developed by Perlovsky (2009b) on the basis of the study by Guttfreund et al. (1990) and several others, different languages are associated with different levels of emotionality. Cross-cultural research provides evidence that people speaking different languages have a different perception of emotion from one another (Lindquist & al., 2015b). Moreover, expressing one's emotions in a second language could be a form of emotion regulation (Rozensky & Gomez, 1983 as in Guttfreund, 1990).

Thus, speaking in a second language should not only influence the participants' emotionality but

also the strategy they use to regulate these emotions if the second language differs in emotionality from the mother tongue. It is unclear, however, in which direction this form of emotion regulation goes (Perlovsky, 2009b). Moreover, the behavior of people expressing their emotions in a second language that Rozensky and Gomez (1983) observed could be interpreted either as an adaptive or a maladaptive form of emotion regulation. Since Richards and Gross (2000) have shown that suppression, a more maladaptive emotion regulation strategy, if used in the long term, can impair cognitive processes such as incidental memory while cognitive reappraisal, a more adaptive strategy, does not have this detrimental effect, the expression of emotions in a second language could also impact cognitive processes. Should the expression of emotion in a second language lead to negative emotion regulation strategies then it would follow that expressing one's emotions in a second language would lead to poorer memory. Should this expression be an adaptive form of emotion regulation then this should not be the case. From the aforementioned theories we derive our exploratory hypothesis that those who express their emotions in their mother tongue will have a different performance on a subsequent cognitive task involving working memory than those who express them in a second language if the second language differs in emotionality from the mother tongue.

Methodology

Material and measures

Socio-demographic data. General demographic data of the participants, including questions about their age, gender, highest level of education, subjective socio-economic status (SSES) and their English proficiency was collected via a questionnaire formulated specifically for this research project's purposes. The subjective socio-economic status was assessed by letting participants rate their family's financial situation by comparing it to society on a scale from one to seven, with seven representing families having the best financial situation and one representing those having the worst. The language classification system used was adopted from the Common European Framework of Reference for Languages (Language Policy Division, Council of Europe, 2001), whereas the descriptions were shortened.

Mental and Emotional state. The current mental state of the participants was assessed using the Positive and Negative Affect Schedule (Watson & Clark, 1988). The PANAS consist of 20 items, adhering either to the Negative Affect or Positive Affect scale. Participants are asked to indicate to what extent they currently feel a certain way using a five-point Likert scale (Likert, 1932) ranging from "not at all" over "a little", "moderately", "quite a bit" up to "extremely". The Positive Affect scale represents the range to which a person feels alert, enthusiastic, and active. If its score is high, the participant feels energetic, a pleasurable engagement, and concentrated, while a low score is defined by lethargy and sadness. The Negative Affect scale constitutes unpleasurable engagement and subjective distress. The higher that score, the more the participant feels anger, contempt, guilt, fear, disgust, and nervousness. A low score indicates a state of serenity and calmness. According to Watson and Clark (1988) this instrument is highly internally consistent as the Cronbach's coefficient α ranges from .86 to .90 for the Positive Affect and from .84 to .87 for the Negative Affect. Time instructions do not influence the scales' reliability. There is a low correlation from -.12 to -.23 between the Positive Affect and Negative Affect, which designates a quasi-independence. The test-retest reliability was measured by conducting the PANAS twice for each different time frame. The correlations between the two administrations for all the time frames were between .39 and .71. The retest stability tends to increase as the rated time frame lengthens. A two-tailed *t* test with $p > .05$ showed no significant differences between the Positive Affect and Negative Affect values regarding their retest stability. Similar results were found with other samples. Finally, the validity of the PANAS is also good as Watson and Clark demonstrate in their paper (1988).

Frustration Task. A 35-piece all white puzzle was used to manipulate participants and generate a negative emotion response, such as frustration. Participants were told that the puzzle was designed for children over the age of eight and that most participants completed the puzzle in three minutes and 12 seconds. The time limit to finish the task was set at five minutes. Many of the puzzle parts could be placed at some positions, but they would mostly clamp. In the beginning, participants were explained that only the one location where the part would fit perfectly was correct. This material is inspired by a similar method used in the study of Rosenzweig (1943), where participants had to solve puzzles but were interrupted by the time they only finished half of it.

Memory Task. A Memory card game with 36 forest motif pairs, was split in two halves and used as assessment of working memory and proxy for cognitive performance. The first half was shown in the pretest and the second in the posttest. Working memory was operationalized as the number of matching pairs that the participants could find within a time limit of two and a half minutes, after 30 seconds of memorization.
Language Manipulation Task. Participants were asked to report on the feelings and thoughts they may have experienced during the puzzle task and the overall study, using a semi- structured interview formulated for this study's purposes. Participants were interviewed in English or in German depending on whether they were in the experimental group or in the control group. The interview consisted of five main questions focusing on the participants' feelings and emotions during the puzzle task and their overall experience of the study. The interviewers monitored the length of responses and kept them about five minutes long. Participants' answers were noted, but the aim of the interview was to let participants talk about their feelings and sensations in a specific language.
Emotion Regulation Strategies. To interrogate the participants about the emotion regulation strategies they potentially applied during the Frustration Task, interviewers conducted a semi- structured interview in German. Participants were asked to explain what they felt, whether they tried to change the way they felt, and how they did so. Two people in our team then interpreted these explanations to determine whether the participants had used any emotion regulation strategies and whether those strategies were maladaptive or adaptive. The Cohen's kappa coefficient was subsequently computed to measure the interrater reliability of these two assessments. It revealed that our raters were in moderate agreement as to whether the participants regulated their emotions $\kappa =$

.507, $p < .001$ and that they did not agree at all on the type of emotional regulation strategy used, $\kappa = -.097$, $p =$
.633. One of the raters' assessments was then chosen randomly.
Emotion intensity. In addition, a question about the intensity of the emotion felt during the frustration task was administered. Participants had to indicate on a six-point Likert scale (Likert, 1932) whether the emotion they felt was "very weak", "weak", "moderate", "strong", or "very strong". They also had the option to check "no emotion" to indicate that the frustration task didn't elicit an emotional reaction in them.
Emotion Regulation Trait. The Fragebogen zur Erhebung der Emotionsregulation bei Erwachsenen (Grob & Horowitz, 2014), an emotion regulation trait questionnaire, was used to assess participants' long-term emotion regulation abilities and habits. *Problem-oriented-action, lightening of the mood, acceptance, forgetting, cognitive problem-solving* and *reappraisal* are considered adaptive strategies. *Giving up, catastrophizing, withdrawing, self-devaluation, blaming others* and *perseveration* make up the maladaptive strategies (Lange & Tröster, 2015). The scores in each of these emotion regulation strategies are then aggregated into two measures: a score for the general usage of adaptive strategies and a score for the general usage of maladaptive strategies. In our study we only focused on whether the emotion regulation strategies that the participants generally use are adaptive or maladaptive and not on the specific strategies because identifying specific emotion regulation strategies during the interview would have been too difficult and not objective enough. According to Grob and Horowitz (2014) the FEEL-E's objectivity is ensured by the standardized instruction, the given response scheme, the profile sheet and the conversion of the raw values into *t*- values. The acceptable Cronbach's coefficient α lies between .73 and .89 for the primary scales, between .65 and .81 for the emotion-specific secondary scales, for the maladaptive strategies at

.88 and for the adaptive strategies at .91. The test-retest reliabilities indicate for the emotion-overlapping scales adequate stability ($.61 < r_n <$ .79) and for the emotion-specific scales acceptable stability ($.61 < r_n < .73$). The intercorrelations within the 12 strategies and the two secondary scales show independence and prove discriminative validity. However, there are positive correlations ($.42 < r < .63$, $p < .001$) within the adaptive strategies, whereas the maladaptive strategies show no noteworthy correlations.

Procedure

To recruit participants, students were approached on the campus of the University of Luxembourg and surrounding areas such as malls and were given a flyer and a brief description about our study. Furthermore, flyers were hung up on different walls at the university. Social media was also used as an advertising tool. At first, the participants received a short cover story regarding the aim of the study to prevent any kind of influence to their behavior and

emotion before the real manipulation. The researchers pretended that the study's aim was to investigate only bilingualism and cognitive performance without mentioning the fact that emotions would also be investigated. Participants were told that our intention was to investigate whether bilinguals have an advantage in visual-spatial reasoning, problem solving and working memory, compared to monolinguals. Then they were truthfully informed about the type of tests they were going to do, their rights, which data would be collected, confidentiality and data protection. Pseudonymity was given by design and data is protected and treated according to the law, the European General Data Protection Regulation and the Ethic's guidelines from the University of Luxembourg (UL). All data is stored in safe archives at the UL and will be destroyed ten years after the publication of this paper.

Participants were invited to the University of Luxembourg and a private apartment in Trier in Germany for an in-person testing session of about 50 minutes. They were asked not to consume caffeine or theine eight hours before the study, to assure that their cognitive activity isn't influenced by any extraneous factors. They sat down on a table across from the tester. They were given the paper-pencil questionnaires for socio-demographic data and their current emotional and mental state, the PANAS.

Subsequently, participants' working memory was assessed for a first time, using the Memory Task, in which they had to find as many matching pairs as possible within the limited amount of time. In order to analyze the possible influence of the experience of a negative emotion and the subsequent (verbal) emotion regulation may have on cognitive abilities (i.e. working memory), this study was designed as a pre-/posttest comparison. Like this, the study design allowed for a direct comparison between cognitive abilities before, and after the Frustration and Language Manipulation Tasks.

After the first Memory Task, the participants were instructed to solve the white puzzle within the time limit, which of course was not possible.

Next, participants were asked to complete the PANAS again and the tester started the semi-structured interview of the Language Manipulation Task. Participants were either interviewed in their mother tongue German, if they belonged to the control group, or in English, if they belonged to the experimental group. After the verbal expression of emotion in one or the other language, participants were asked to complete the other half of the working memory task.

To control for the influence that participants' emotion regulation might have had on their working memory, participants were asked to respond to a second semi-structured interview on emotion regulation strategies. Subsequently they were asked to rate the intensity of the evoked emotions, to account for the effects that the intensity of emotions might have had on working memory. Finally, the FEEL-E questionnaire was used to measure whether the participants generally prefer maladaptive or adaptive regulation strategies.

To debrief the participants, they received a full explanation about the real purpose of the study and why they needed to be deceived in the beginning. They now have been told that the interest was to find out if the language, in which we express our emotions, influences cognitive performance. They have also been enlightened about the use of the tests and why the puzzle was unsolvable. By ensuring the participants that the failure of the puzzle isn't his or her fault, any long-term damage, confusion or diminished self-confidence has been prevented. Finally, participants were also adequately compensated for their participation by receiving a ten euros SODEXO voucher, which can be redeemed in many luxembourgish stores.

Sample

The sample we recruited consisted of N = 29 participants whose mother tongue was German and their second language English. People with mental illnesses or any disorders affecting cognition, as well as people who take stimulants or medication which affects cognition, were excluded from the study. 51.72% of the participants were assigned to the control group and 48.28% to the experimental group. This assignment was done randomly. Participants ranged in age from 20 to 61 years although the majority of the sample contained people in their 20's. The mean age was therefore M = 25.21 (SD = 8.845) years. Moreover, 55.2% of the participants were female and 44.8% male. The most common highest level of education was high school (75.9%) which can be explained by the fact that most of our participants were University students. Additionally, the subjective socioeconomic status ranged from three to six, with five being the most common value (44.8%), and four the second most common (34.5%). Finally, the participants' English proficiency ranged from B1 to C2, where C1 represented the most common level (48.3%).

Analyses

First, we conducted preliminary analyses to check the assumptions of the tests and models we used, to look for outliers and to identify other problems which would interfere with the interpretation of our results. Hence Mann-Whitney *U* tests and a chi- squared test were conducted to check whether gender, age, education and subjective socioeconomic status varied significantly between our control and our experimental group. We also identified a participant whose difference in working memory performance constituted an outlier by looking at the boxplot of this variable. As a result, this participant was excluded in all analyses involving the difference in working memory performance. Afterwards we verified whether our puzzle task had elicited a negative emotion, such as frustration, by using a two-tailed paired sample *t*-test with which we compared negative and positive affect measured by the PANAS before and after the frustration task within the two groups. In order to analyze the possible influence that the experience of a negative emotion and the subsequent verbal emotion regulation may have on cognitive abilities (i.e. working memory), this study was designed as a pre-/posttest comparison. This study design allowed for a direct comparison between cognitive abilities before and after the Frustration and Language Manipulation Tasks. Therefore, to check whether the difference in pre- and post-test working memory performance within the two groups was significant, a two-tailed paired sample t-test was conducted. This difference in pre- and post-test working memory performance was computed by subtracting the working memory performance during the pre-test from the performance during the post- test. We then compared the difference in working memory performance pre- and post-test between our control- and experimental group with a two-tailed

independent sample t-test to analyze whether the expression of emotion in different languages had any effect. Moreover, correlation analysis between the affect measured by the first PANAS administration and the pre-test working memory performance and between the affect measured by the second PANAS administration and the post-test working memory performance were run to account for the influence of the participants mood on their performance. To control for extraneous variables, we examined the relationship between the difference in working memory performance and the intensity of emotions, age, gender, education and subjective socioeconomic status. Furthermore, the English language proficiency of the experimental group was correlated with their difference in working memory performance to examine whether the English language proficiency might affect the relationship between the difference in working memory performance and the expression of emotion in different languages. Finally, we built a multiple regression model with the backward method using the type of emotional regulation strategy used during the puzzle task and the type of emotional regulation strategy that the participants use the most in general as predictors for the difference in working memory performance. This analysis was conducted to account for the fact that different emotional regulation strategies have different effects on cognitive performance and to check whether our results might have significantly been influenced by this relationship.

Results

Preliminary analyses. The results revealed that there are no significant differences between our control group and our experimental group in terms of age $U = 101.5$, $p = .88$, subjective socioeconomic status $U = 88.5$, $p = .477$, education, $U = 71$, $p = .146$, and gender χ^2 (1, $N = 29$) = .293, $p = .588$. The mean

and range of these variables are summarized in Table 1. A Mann- Whitney *U* test was used instead of an independent samples t-test to analyze the differences in terms of age since a Shapiro-Wilk test revealed that the distribution of the variable age, $W(29) = .605$, $p < 0.001$, did not follow a normal distribution, meaning that the assumptions for the independent sample t-test were not fulfilled.

Table 1: Median and range for the variables age, education, subjective socio-economic status and gender

	median	range
age	23	41
education	4	4
Subjective socio-economic status	5	3
gender	1	1

Note. The median and range were reported instead of the mean and SD since these values were used to detail the results of a non parametric test, the Mann-Whitney U test.

Verification of the effectiveness of our emotion manipulation. Furthermore, our results showed that our puzzle task was successful in eliciting negative emotion since the negative affect of the participants was significantly higher (M = 1.86, SD = .559) after the puzzle task than before it (M = 1.29, SD = .269), $t(28) = 6.761$, $p < .001$. Similarly, positive affect was significantly lower (M = 2.66, SD = .717) after the puzzle task than before it (M = 2.93, SD = .499), $t(28) = -2.353$, $p < .05$. These results are illustrated in Figure 1.

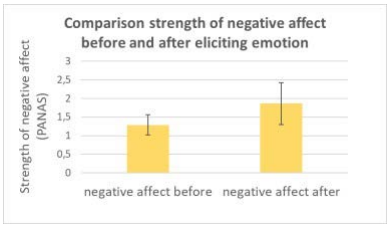


Figure 1: Negative affect of the participants before and after the puzzle task

Differences in working memory performance between groups and between the pre- and post-test. Moreover, there were significant differences in participants' working memory performance between the pre- and the post-test with participants finding significantly more pairs during the post-test (M = 9.79, SD = 4.095) than during the pre-test (M = 7.96, SD = 4.238), $t(27) = 2.862$, $p < .01$. However, there were no significant differences between the control group (M = 1.53, SD= 3.701) and the experimental group (M = 2.15, SD = 3.051) when it came to the difference in their working memory performance, $t(26) = -.479$, $p = .636$.
Therefore, the expression of emotion in different languages did not affect working memory performance in our sample. These results are illustrated in Figure 2 and 3 respectively.

Figure 2: Difference in working memory performance between the pre- and post-test

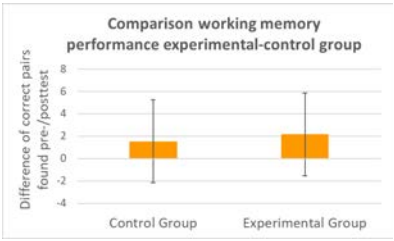


Figure 3: Comparison of the difference in working memory performance between the control group and the experimental group

Analysis of the potential influence of mood and English language proficiency on working memory performance. Correlation analyses showed that the mood of the participants before the memory task did not affect their working memory performance during the pre-test and that their mood after the frustration task did not affect their performance during the post-test. Correlations between working memory performance during the pre-test and the first PANAS administration were $r = -.185$, $p = .337$, and $r = .159$, $p = .411$, for the negative and the positive affect respectively. As for working memory performance during the post-test and the second PANAS administration, they were $r = -.049$, $p = .799$, and $r = -.072$, $p = .71$. Additionally, there was a significant relationship between English proficiency and the difference in working memory performance in the experimental group, $r_s = -.532$, $p = .05$.

Multiple regression. As for the multiple regression model, only the emotion regulation used during the puzzle task and the emotion regulation used by the participants in general were used as predictors for the difference in working memory performance since correlation analyses revealed that there were no significant relationships between the difference in working memory performance and age, gender, education, subjective socioeconomic status and the intensity of emotions. These correlations are shown in Table 2. However, neither emotion regulation as a trait $\beta = -.481$, $p = .829$, nor the type of strategy used during the frustration task, $\beta = -.897$, $p = .719$, could significantly predict the difference in working memory performance when the enter method was used $F(2, 17) = .077$, $p = .926$, $R^2 = .095$. A significant effect could not be found either with the backward method: in this case the type of emotion regulation strategy used during the frustration task alone, $\beta = -.784$, $p = .741$, could still not significantly predict the difference in working memory performance $F(1, 17) = .113$, $p = .741$, $R^2 = .079$. Table 3 provides a summary of the multiple regression analysis.

Table 2: Correlations between various extraneous variables and the difference in working memory performance

	Difference in working memory performance
Intensity of emotions	r = .088, p = .651
age	r = .038, p = .848
gender	r = .137, p = .487
education	r_s = -.084, p = .671
Subjective socio-economic status	r_s = -.21, p = .284

Table 3: Summary of multiple regression analysis for variables predicting the difference in working memory performance

	Model 1				Model 2			
Independent variable	B	SE	t	p	B	SE	t	p
Emotion regulation as a trait	-.481	2.191	-.219	.829				
Emotion regulation strategy used during the puzzle task	-.897	2.455	-.366	.719	-.784	2.336	-.336	.741
Constant	3.147	3.114	1.011	.326	2.667	2.154	1.238	.232
	Model 1				Model 2			
R^2	= .009				= .006			
F-ratio	= .077	p = .926			= .113	p = .741		
SEE	= 3.833				= 3.73			
n	= 29							

Note. The first model was computed using the entry method and the second model by using the backward method.

Discussion

General Discussion

The aim of the study was to test whether the expression of emotions have a different influence on the performance in a subsequent task that involves working memory depending on whether they are expressed in the mother tongue or in a second language. The hypothesis is founded on Lindquist et al.'s (2015b) findings that the language we speak influences our emotionality. Another key theory was the Dual Model of Language and Cognition by Perlovsky (2009a) which states that emotion regulation capabilities should be differently developed in the mother tongue. Moreover, according to Richards and Gross (2000), emotion regulation influences cognitive performance. Therefore, our hypothesis is that those who express their emotions in their mother tongue will have a different performance on a subsequent cognitive task involving working memory than those who express them in a second language.

No significant demographic differences between the control and experimental group could be found. This is ideal since Lowe et al.'s (2021) meta-analysis mostly found a significant effect of bilingualism on cognitive performance, such as executive functioning, because socio-economic status acted as an extraneous variable. Therefore, it can be said that our results were not influenced by the measured group differences between the control and experimental group.

The emotion manipulation task was successful since it managed to elicit negative emotions such as frustration in almost every participant. Analyses could confirm that the negative affect of the participants was significantly higher after the puzzle task than before it. These are excellent results since the continuation of the study and the interpretation of all results regarding cognitive performance would be meaningless if no emotion had been elicited. As for cognitive performance, the study could find ignificant differences between pre- and post-test working memory performance with participants finding significantly more pairs during the post-test than during the pre-test. This is surprising as one would assume, based on the findings of Gross (2000), that cognitive performance should be poorer or at most equal after a negative emotion has been elicited since the regulation of emotions often uses up cognitive resources. Even if it doesn't use up cognitive resources, currently no emotion regulation strategy is known that improves cognitive performance from its baseline. It is therefore very likely that these findings could simply be attributed to a learning effect.

However, there were no significant differences between the control group and the experimental group when it came to the difference in their working memory performance between the pre- and post-test. Therefore, the expression of emotion in different languages did not seem to have affected working memory performance in our sample. There are multiple explanations for this. It may be that the expression of emotions in different languages simply does not affect cognitive performance. Then again there might indeed be a link between the expression of emotions in different languages and cognitive performance, but with the emotionality of the language acting as a key variable, as the studies by Guttfreund (1990), Rozensky and Gomez (1983), Perlovsky (2009b), Lindquist et al. (2015b) suggest. In that case, German and English may not vary significantly in their emotionality which could explain why no significant differences in working memory performance between the control and experimental group were found. Finally, the limitations of this study could also explain why no significant relationship between the two groups could be found.

A multiple regression analysis revealed that neither the emotion regulation used during the puzzle task nor the emotion regulation used by the

participants in general could significantly predict the difference in working memory performance. Gross (2000) has discovered that participants who use a maladaptive emotion regulation strategy, such as suppression, should perform worse on some subsequent tasks involving memory such as tasks requiring the use of incidental or verbally encoded memory. Therefore, these results seem unexpected at first. But since Gross (2000) could not observe the same detrimental effect of suppression on non-verbal memory they might actually be consistent with the literature since nonverbal memory was tested in this study by using a memory game. It is also possible that participants were not aware of using their second language English as an emotion regulation strategy during the emotion regulation interview since it is a rather unconscious process. The intensity of emotions was not used as a predictor in the multiple regression model since correlation analyses revealed that there was no significant relationship between this variable and the difference in working memory performance. A study that Richards and Gross conducted in 1999 also found that the intensity of emotions and the degree of memory impairment were independent even if they were referring to incidental memory and not non-verbal memory.

We only found a negative relationship between English proficiency and the pre-post-test difference in working memory in the experimental group. The better people reported to speak English, the smaller the pre-post-test difference in working memory performance. This might be explained by a higher level of English proficiency leading to a higher level of emotion regulation competence as language and cognition evolve together (Perlovsky, 2009a). Those who spoke English almost as well as their mother tongue and most likely on a daily basis have a different emotional connection to the second language caused by their experiences while using it. Thus, emancipatory detachment (Kellman, 2000) might have played a bigger role for those with lower English proficiency levels who probably spoke English less often. They regulated their emotion by merely speaking English without needing further strategies.

Including only lower levels of English proficiency using a bigger sample might be beneficial to more significant study results. Even though we found no significant differences in the pre-post-test difference in working memory performance between experimental and control group, the effect of English proficiency on the difference in working memory performance before and after emotion expression leads to the conclusion that expressing one's emotions in a second language might influence cognitive performance.

In sum, the hypothesis that those who express their emotions in their mother tongue will have a different performance on a subsequent cognitive task involving working memory than those who express them in a second language could not be supported because there was no significant difference found between the experimental and control group regarding the cognitive performance. The emotion manipulation task was successful: it managed to elicit negative emotions such as frustration in almost every participant which was necessary before testing the hypothesis. In concordance with Gross' (2000) research, a multiple regression analysis revealed that neither the emotion regulation used during the puzzle task nor the emotion regulation used by the participants in general could significantly predict the difference in working memory performance for our non-verbal task. As participants may not have been able to consciously report using their second language as an emotion regulation strategy, we cannot entirely exclude this explanation in our considerations. We only found a negative relationship between English proficiency and the pre-post-test difference in working memory in the experimental group which might have been caused by a lower level of emancipatory detachment (Kellman, 2000) in those who reported to speak English on a higher level. As emancipatory detachment (Kellman, 2000) was a key variable to our explanation, a lower level of English proficiency might be beneficial to more significant study results. This effect also shows that expressing one's emotions in a second language might influence cognitive performance even though we found no significant differences between experimental and control group.

Limitations

The fact that the participants found more pairs in the post-test than during the pre- test of the working memory task could be due to the fact that there was a learning effect. The reason for the participants' improvement in performance could therefore simply be attributed to them getting used to the memory game and would in this case not be due to differences in emotion expression. Moreover, some participants are more experienced in playing memory games than others, either because they have siblings or because they like these kinds of games. Moreover, the time it took to turn over the memory cards by hand was not always the same which could have led to better memory in those participants who had more time to memorize the cards.

Regarding the criteria for participation in the study, we could only test participants aged 20

years and older, as this criterion had to be fulfilled for the FEEL-E. This already excluded many test participants, since students in the bachelor's programs are often younger. Not to mention that time available to recruit the participants was too short to gather enough participants. The small sample size was also due to the current pandemic situation and its restrictions. It was also difficult to find many people with a German mother tongue in Luxembourg. Due to our recruitment difficulties, we also started testing in Trier. As a result, the testing conditions were not standardized anymore. The room at the university was an artificial setting without environmental confounding variables. This was not possible in Trier, so we cannot exclude bias due to interfering variables. In general, the number of participants influences the fit of the regression model which might be better with more participants. The sample size may also be too small to obtain statistically significant effects between the experimental and the control group and to generalize it to the whole population. Moreover, while indicating their English proficiency, many participants stated that they could understand nearly everything in English but that they had not spoken English since their school days. Therefore, the self-reported English proficiency might not have been reliable. This should be taken into account when considering the negative relationship between English proficiency and thepre-post-test difference in working memory performance that we found in the experimental group. During the interview, some people also had difficulties because of their limited English vocabulary, and it influenced the expression of emotions and feelings. This may have distorted the study results and made them less valid.

Another flaw of our measures was that we did not specify which type of emotions the participants should rate when we asked them to rate the intensity of their emotions elicited by the puzzle. Moreover, it was complicated to interpret whether the participants had used an emotion regulation strategy and whether that strategy was adaptive or maladaptive, since this type of rating is highly subjective. This explains the low inter reliability rating of the emotion regulation strategies interview. Furthermore, some participants seemed to have used more than one strategy. This could explain why the multiple regression model could not find a significant relationship between the emotion regulation used and the difference in working memory performance. It must also be noted that an emotion regulation strategy that is maladaptive when used on a regular basis can be adaptive, depending on the context, if it is only used once. Finally, many participants reported that they had never thought about their emotion regulation strategy and that they were thus not used to talking about it. Emotion regulation is also not a topic about which one talks openly with everyone. These factors together with social desirability might have resulted in the participants not answering the questions about emotion regulation accurately or honestly.

Outlook

Future studies could improve our study by increasing the sample size and by recruiting a more representative sample that does not mostly consist of participants in their twenties. They could also use a standardized emotion eliciting task, where it is clear that frustration or rather one specific emotion is elicited. A standardized test is not a guarantee but may improve the chances that most participants will have a similar emotion. In this study it could have happened that we elicited another emotion than frustration because we did not test if it was for sure frustration or another emotion. They could also differentiate between positive and negative emotions and the effect of emotion expression of positive/negative emotions in different languages on cognitive performance. The difference between the participants due to the researcher turning over the memory cards with different timings could be improved by already preparing a covered game and not having a strong bias due to the experimenter turning over the cards for different periods of time. In further studies, the performance of the memory task could be made more objective by integrating technical tools like a computer to minimize the biases by the experimenter. In future studies it would be important to test the participants' English proficiency beforehand with a standardized language test. This would be done to be sure that the participants do not have problems with their articulation in the interviews or difficulties regulating their feelings in the asked language because of the minimal language proficiency. Future studies could also conduct a state emotion regulation questionnaire instead of an interview to allow for an objective interpretation of the used emotion regulation strategies. So far, there is only one standardized questionnaire in English namely the State Emotion Regulation Inventory (Katz & al., 2017) but simply translating it into German would not assure that the quality criteria of the questionnaire remain unchanged.

Conclusion

All in all, there was no significant relationship

between the expression of emotion in different languages between the groups and the difference between pre- and post-test in working memory performance even though our analyses confirmed an increase in negative affect after the puzzle task. That led to the fact that we could not confirm the hypothesis. Neither emotion regulation as a trait, nor the type of strategy used during the frustration task could significantly predict the difference in working memory performance. Since a higher English proficiency seems to have led to a smaller difference in working memory performance between the pre- and post-test, we might have found a difference in working memory performance between the control and experimental group if we had recruited participants with a lower English proficiency. According to our current knowledge and considering the time limit, the study has been successful overall. There are many suggestions to minimize the limitations and a lot of different possible opportunities of improvement that can be extended in further studies.

Funding: The present study was supported by the Luxembourg National Research Fund (FNR) (13651499).

References

Bialystok, E. (2011). Coordination of executive functions in monolingual and bilingual children. *Journal of Experimental Child Psychology, 110*(3), 461-468.https://doi.org/10.1016/j.jecp.2011.0 5.005

Bialystok, E., Craik, F. I. M., & Luk, G. (2012). Bilingualism: Consequences for mind and brain. *Trends in Cognitive Sciences, 16*(4), 240–250. https://doi.org/10.1016/j.tics.2012.03.001

Bialystok, E. (2017). The bilingual adaptation: How minds accommodate experience. *Psychological Bulletin, 143*(3), 233–262. https://doi.org/10.1037/bul0000099 Festinger, L., & Carlsmith, J. M. (1959).

Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology, 58*(2), 203–210. https://doi.org/10.1037/h0041593 Franklin, A., Drivonikou, G. V., Bevis,

L., Davie, I. R. L., Kay, P., & Regier, T. (2008). Categorical perception of color is lateralized to the right hemisphere in infants, but to the left hemisphere in adults. *PNAS, 105*(9), 3221–3225. https://doi.org/10.1073/pnas.0712286105

Grob, A., Horowitz, D. (2014). Fragebogen zur Erhebung der Emotionsregulation bei Erwachsenen. *Hans Huber, Hogrefe AG*.

Gross, J. J. (1998). The Emerging Field of Emotion Regulation: An Integrative Review. *Review of General Psychology*, *2*(3), 271–299. https://doi.org/10.1037/1089-2680.2.3.271

Guttfreund, D. G. (1990). Effects of language usage on the emotional experience of Spanish-English and English-Spanish bilinguals. *Journal of Consulting and Clinical Psychology, 58*(5), 604–607. https://doi.org/10.1037/0022-006X.58.5.604

Izard C. (2010). The Many Meanings/Aspects of Emotion: Definitions, Functions, Activation, and Regulation. *Emotion Review, 2*(4), 363-370. https://doi.org/10.1177/1754073910374661

Katz, B. A., Lustig, N., Assis, Y., & Yovel, I. (2017). Measuring regulation in the here and now: The development and validation of the State Emotion Regulation Inventory (SERI). *Psychological Assessment, 29*(10), 1235–1248. https://doi.org/10.1037/pas0000420

Kellman, S. G. (2000). *The translingual imagination* (1st ed.). University of Nebraska Press. https://books.google.de/books?hl=en&lr=lang_en|lang_fr|lang_es|lang_de&id=2jp0xviQY9IC&oi=fnd&pg=PR7&dq=kellman+2000+the+translingual+imagination&ots=kNPsbb34AQ&sig=fG2yry2CQLI7R0gjU9FnV0dKt2w#v=onepage&q=kellman%202000%20the%20translingual%20imagination&f=false

Kroll, J. F., & Bialystok, E. (2013). Understanding the consequences of bilingualism for language processing and cognition. *Journal of Cognitive Psychology, 25*(5), 497–514. https://doi.org/10.1080/20445911.2013.799170

Lange, S., Tröster, H. (2015). Adaptive und Maladaptive Emotionsregulationsstrategien im Jugendalter. *Zeitschrift für Gesundheitspsychologie, 23*(3), 101-111. doi:10.1026/0943-8149/a000141

Language Policy Division, Council of Europe. (2001). Common European Framework of Reference for Languages: Learning, teaching, assessment. *Cambridge University Press*. https://rm.coe.int/1680459f97

Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archive of Psychology*.

Lindquist, K. A., MacCormack, J. K., & Shablack, H. (2015a). The role of language in emotion: predictions from psychological constructionism. *Front. Psychol. 6:444*. doi: 10.3389/fpsyg.2015.00444

Lindquist, K. A., Satpute, A. B., Gendron, M. (2015b). Does Language Do More Than Communicate Emotion?. *Current Directions In Psychological Science, 24*(2), 99-108. https://doi.org/10.1177/0963721414553440

Lowe, C. J., Cho, I., Goldsmith, S. F., & Morton, J. B. (2021). The Bilingual Advantage in Children's Executive Functioning Is Not Related to Language Status: A Meta-Analytic Review. *Psychological Science, 32*(7), 1115–1146. https://doi.org/10.1177/0956797621993108

Ochsner, K. N., & Gross, J. J. (2008). Cognitive Emotion Regulation. *Current Directions in Psychological Science,17*(2), 153–158. doi:10.1111/j.1467-8721.2008.0056

Perlovsky, L. (2009a). Language and Cognition. *Neural Networks, 22*(3), 247–257. https://doi.org/10.1016/j.neunet.2009.03.007

Perlovsky, L. (2009b). Language and emotions: Emotional Sapir-Whorf hypothesis. *Neural Networks, 22*(5– 6), 518–526. https://doi.org/10.1016/j.neunet.2009.06.034

Ramírez-Esparza, N., Gosling, S. D., Benet-Martínez, V., Potter, J. P., Pennebaker, J. P. (2006). Do bilinguals have two personalities? A special case of cultural frame switching. *Journal of research in Personality, 40*(2), 99-120. https://doi.org/10.1016/j.jrp.2004.09. 001

Richards, J. M., & Gross, J. J. (1999). Composure at Any Cost? The Cognitive Consequences of Emotion Suppression. *Personality and Social Psychology Bulletin, 25*(8), 1033–1044. https://doi.org/10.1177/0146167299 2511010

Richards, J. M., & Gross, J. J. (2000). Emotion Regulation and Memory: The Cognitive Costs of Keeping One's Cool. *Journal of Personality and Social Psychology, 79*(3), 410–424. https://doi.org/10.1037/0022-3514.79.3.410

Rosenzweig, S. (1943). An experimental study of 'repression' with special reference to need- persistive and ego-defensive reactions to frustration. *Journal of Experimental Psychology, 32*(1), 64–74. https://doi.org/10.1037/h0062000 Rozensky, R. H., & Gomez, M. Y. (1983). Language switching in psychotherapy with bilinguals: Two problems, two models, and case examples. *Psychotherapy: Theory, Research & Practice, 20*(2), 152–160. https://doi.org/10.1037/h0088486

Watson, D., & Clark, A. C. (1988). Positive and Negative Affect Schedule. *Journal of Personality and Social Psychology, Vol 54*(6). American Psychological Association. https://psycnet.apa.org/buy/1988-31508-001

Whorf, B. L. (1956). Language, Thought and Reality. *MIT Press*. https://psycnet.apa.org/record/1956-07134-000

Is the early face processing disrupted by medical face masks?

Emilie Backes, Charlotte Gebhardt, Hannah Mareike May, Leila Muhovic, Nidara Rahic and Sirinda Tintinger

Supervisor: Dr. Annika Lutz

Due to the implementation of mandatory masking in an effort to minimize the impact of the COVID-19 pandemic, the extent to which facial masks change the way faces are perceived is analyzed. Facial features and their relationship to each other have a significant impact on configural processing of the face and emotion recognition, therefore this study examines the extent to which masking the lower half of the face may influence emotion recognition, configural and emotion processing. Previous research has shown specific results regarding the N170 as well as other studies relate to the impact of masks. Our study focusses on: “Is the early face processing disrupted by medical face masks? “. Thirty participants were recruited for this study which took place under the COVID-19 restrictions at that time, including 65.50% females and 34.50% males. Our results showed that the N170 was more pronounced once the medical face mask was added to the photographs. Moreover, the inversion effect in the interaction effect between mask and orientation on the N170 regarding the unmasked faces was stronger, which indicates that configural processing is made more difficult by the medical face mask. Regarding emotion recognition, the mask showed a significant effect, as did the interaction between the orientation and the mask. Furthermore, the interaction between the emotions, the mask, and the orientation showed that emotion recognition was made more difficult under the condition that the medical face mask was added to an inverted picture. A significant effect was observed exclusively for the emotions fear, happiness, and neutral. Considering previous studies, we could show that wearing masks impeded emotion recognition and configural processing, but no relevant effects on emotion processing were indicated. However, configural processing was shown to be impeded by the wearing of masks.

Introduction

There has been a large scientific interest in faces, their processing and their importance in social interactions for a number of years. Faces seem to be favoured from the moment of birth, and this continues over the whole life (Pascalis, 2002; Bloomington, 1996). This may be due to the fact that the face provides important information to humans as social creatures and also enables them to share their own impressions through non-verbal communication. Faces reveal important information about a person's identity, support speech analysis considerably and give information about how a person is feeling and much more (Carbon, 2020).

Depending on the literature, different face processing methods are proposed, with the majority distinguishing between configural, featural and holistic processing. In featural processing of the face, the focus is on the individual facial features, i.e. eyes, nose and mouth and their details, such as size, shape or colour (Schmid et al., 2013; Carbon, 2011; Freire et al., 2000; Derntl & Carbon, 2009; Itier & Taylor, 2004). Configural processing of faces is often further subdivided (Freire et al., 2000; Derntl &

Carbon, 2009). First-order configural processing refers to the spatial relationship between the features just mentioned, which means: the eyes are above the nose, which itself is above the mouth (Derntl & Carbon, 2009; Schmid et al., 2013 ; Carbon et al., 2005). Second-order configural processing refers to the spatial relationship of the features, i.e. the distances between eyes, nose and mouth (Schmid et al., 2013, Hole et al., 1999; Carbon, 2011; Itier & Taylor, 2004). Finally, holistic processing is the processing of the whole face, which is not separable into divided features (Carbon et al. 2005 ; Itier & Taylor, 2004; Schmid et al., 2013; Derntl & Carbon, 2009). Facial researchers also suggest that features and their relationship to each other are processed holistically (Carbon, 2011). This can be seen, for example, in the study by Hole et al. (1999). Here, chimeric faces, i.e. the faces of two people, were put together as one. It was noticeable that the visual separation of the two halves of the face was very difficult as long as the upright face was presented. Only when this was turned upside down the participants were able to distinguish between the two halves of the face. This illustrates that when the face is viewed upright, the spatial relationship prevents the attention to the upper part, whereas configural processing no longer works for inverted faces (Freire et al., 2000; Hole & George, 1999). A similar effect has also been illustrated by the Thatcher Illusion. Here, only the facial features (eyes and mouth) were reversed. While this was immediately noticeable when the faces were presented upright, it was not noticeable when the images of the faces were inverted (Hole et al., 1999).

In order to study the neural processes of face perception, electroencephalography (EEG) is most commonly used, as its millisecond resolution allows examination of neural activity in almost real time (Feuerriegel et al., 2014; Hinojosa et al., 2015). Within EEG measurements, we are particularly interested in event-related potentials (ERP), as these provide us with information about the temporal sequences of brain processes. Among these, we mainly are interested in the negative ERP N170, which reacts more strongly to facial stimuli and is recorded in parieto-occipital areas (Maurer et al., 2007). Furthermore, some articles showed a more pronounced N170 in the right hemisphere (Wright & Kuhn, 2017; Bloomington, 1996; Sträter, 1992). The mentioned face sensitivity is shown by the fact that the N170 amplitude is larger for faces than for objects and non-faces, with a shorter latency (Itier & Taylor, 2004; Eimer & Holmes, 2002; Hinojosa et al., 2015; Pascalis, 2002). This increased amplitude occurs approximately 170 ms after stimulus presentation, which is how the name was derived (Sträter, 1992; Feuerriegel et al., 2014). Another indication that points to the sensitivity to faces of the N170 is the so-called inversion effect. When a face is inverted, this produces a larger or more negative amplitude and a longer latency, although this effect does not occur for inverted objects (Itier & Taylor, 2004; Kloth et al., 2013).

This increased amplitude, however, seems to represent that processing does not occur as intended. There have been different explanations for this effect depending on the researcher. On the one hand, the assumption was made that face-sensitive areas, such as the fusiform gyri, are also activated when the face is inverted and that the inversion of the face additionally activates areas of the cortex that are responsible for processing objects, which leads to an overall increase in amplitude (Pascalis, 2002). Since there was also a comparable increase in amplitude and delayed latency when only the isolated upright eye region was presented, although this effect was absent for other facial features such as the nose and mouth (Bloomington, 1996; Kloth et al., 2013), Itier suggested that this effect was due to the fact that there must be eye-sensitive as well as face-sensitive cells in the occipito-temporal regions of the face-sensitive area (Kloth et al., 2013). More specifically, she suggested that as long as the face is presented the right way around, the N170 is initiated by face-sensitive cells, while the eye-sensitive cells are inhibited. However, this inhibition is lifted as soon as the face is turned around, thus both cells are activated and therefore cause a larger amplitude.

The inversion effect makes clear that the configural processing of the face is more important than the featural. If an image of a face is inverted, processing is more difficult than before

and takes longer, although the features are unchanged and the face as a whole is still present. However, only the spatial relationship is changed, which demonstrates the importance of configuration for the processing of faces (Schmid et al., 2013; Freire et al., 2000; Hole et al., 1999; Itier & Taylor, 2004; Kloth et al., 2013).

Now, for almost two years, we have been in a period of much caution about the SARS-CoV-2 virus that has caused a pandemic and many people to fall ill or even die. This has led more and more countries making face masks mandatory in public places and other places of interpersonal gathering such as restaurants and supermarkets. On the one hand, face masks have taken a very important role at the moment, as they limit the spread of the virus, protect vulnerable people and help to avoid social isolation. However, an increasing number of people is concerned that this could also lead to disadvantages in our social behaviour. As mentioned earlier, we use different types of processing to perceive faces, with configural processing being very important for face perception. With the mask, it can be assumed that the configural processing of faces is significantly disrupted, making it much more difficult for us to process them successfully.

Furthermore, there is also disagreement in the literature about whether the processing of faces is a different, parallel process or a common process when it comes to the processing of identity and emotion (Hinojosa et al., 2015). For example, different processing locations for identity and emotion have also been proposed. The initial visual perception of faces (e.g. identity) is thought to be processed in inferior occipital gyri, while superior temporal sulci are responsible for the perception of changing aspects such as facial expressions (Barborik, 2012; Blais et al., 2012; Wright & Kuhn, 2017). Furthermore, it has long been assumed that N170 is only responsible for the structural processing of faces, such as identity, and that only later ERP could process facial expressions (Maurer et al., 2007). However, this has proven to be false, as the meta-analysis by Hinojosa et al. (2015) showed. Here it was clarified that emotions initiated larger N170 amplitudes than neutral faces. In particular, angry, fearful and happy facial expressions triggered increased N170 amplitudes in descending order. It has been suggested that these are emotions that require rapid responses, which may be why these emotions are processed more intensely (Hinojosa et al., 2015). Taken together, the N170 is thought to represent multiple sources of neural activity (Feuerriegel et al., 2014).

When dealing with the emotion perception, it seems to be very important to identify the relevant areas of the face that can provide helpful information about the emotional state. Again, the literature does not seem to agree on whether the lower or upper half of the face is more important. However, most emphasise that both the eye and mouth regions are similarly important depending on the emotion (Wegrzyn et al. 2017). Eisenbarth & Alpers (2011) further point out that the eyes were fixated at least as often as the mouth for each emotion in their study. However, some researchers attach particular importance to the mouth region, which, according to Blais et al. (2012) and Kotsia et al. (2008), offers more relevant information for the recognition of emotions.

Since masks now cover most of the face and only the eyes are accessible, it can be assumed that emotion recognition is even more difficult and that there is considerably more confusion, which can have serious consequences in our interactions, since emotions can usually be understood universally, causes feelings and often provide deeper insights than verbal communication.

Some studies have already investigated how covering the lower half of the face can affect emotion recognition. In addition, there have also been some studies on emotion recognition with a face mask, all reporting a drop in accuracy in emotion assessment (for example: Grundmann et al, 2021; Carragher, 2020; Carbon, 2020; Freud et al, 2020). Carragher (2020) showed how much masks, for example, affected identity recognition even when the faces were known to the participants. This again points to problems in processing facial information with masks. Furthermore, the effects of masking the lower half of the face also seem to have opposite effects. Fischer et al. (2012) showed in their study that emotions were perceived more negatively in people who

wore a niqab, which in his opinion seems to be mainly due to the fact that the mouth region, and thus happy emotion recognition, seems to be impaired. In contrast, Grundmann et al. (2021) found in their study that negative emotions presented to the participant seemed less negative than when no mask was worn, from which he concludes that there seems to be a kind of "positivity bias" due to the ambiguity of emotions. Calbi et al. (2021) and Bassili seem to agree that anger is more dependent on the eye region, and thus, the recognition of angry emotions is less affected by masking of the lower half of the face, while this masking seems to have mainly negative consequences for happy emotions. However, Carbon (2020) discovered in his study that in addition to a reduction in the accuracy of emotion recognition, the emotions of happiness, disgust and sadness, the emotion anger was also more difficult to recognise with a face mask. Nevertheless, the mask did not seem to have any effect on the emotion recognition of fearful and neutral faces. In contrast, Gulbetekin (2021) also discovered that the mask significantly reduced the recognition of the emotion fear. Still, both discovered that happy faces in particular were mistaken for neutral (Carbon, 2020; Gulbetekin, 2021), with Carbon (2020) also reporting confusion of the emotions anger and sadness with a neutral condition.

According to the previous findings, we would like to investigate in our study what effects the wearing of a face mask has on the processing and recognition of faces as well as their emotional expressions. As already mentioned, the facial features (eyes, nose and mouth) and their relationship to each other have a significant influence on the configural processing of the face. However, since the mask now covers the nose and mouth, the spatial arrangement is disturbed. Accordingly, our first hypothesis is: Configural face processing becomes impaired due to the use of medical face masks. As Hinojosa et al. (2015) have shown in their meta-analysis, some emotional expressions seem to be processed more intensely and thus trigger an increased N170 amplitude. However, this was found for faces that were fully visible. We assume that the mask reduces the processing of emotions when only the upper half of the face is visible. Thus our second hypothesis is: Emotional processing is impeded when looking at faces wearing a medical face mask. Finally, we would like to test in our study to what extent the mask affects the visual recognition of emotions in general. Since emotion recognition is only possible through the eyes, we assume that the recognition of emotions is significantly reduced by the face masks. Accordingly, our final hypothesis is: Recognition of emotional expressions from faces wearing medical masks is impaired.

Methods

Participants

Thirty participants were recruited for this study including both genders (65.50% females and 34.50% males) in an age range of 20 to 52 years (M= 27, SD= 0.50). There was a wide variety of nationalities (56.70% Luxembourgish; 23.30% German; 3.30% French; 16.70% Other). The participants included 37.90% Master graduates; 31% Bachelor graduates; 17.20% High School graduates and 13.90% Vocational trainers. They were voluntarily recruited during the months of July to November 2021. A good knowledge of the German or English language was required and each participant was compensated for their time (35 euros Sodexo gift voucher for 3.5h). People affected by a neurological disease as epilepsy or pregnancy and breastfeeding were excluded from the study. All participants had corrected-to-normal vision and provided informed consent. The study was approved by the Ethics Review Panel of the University of Luxembourg. Furthermore, the study took place under the COVID-19 restrictions at that time (i.e., FFP2 masks had to be worn at all the time during participation).

Materials

All face stimuli were obtained from the MPI FACES database (Ebner, N.C., Riediger, M., & Lindenberger U., 2010). Six frontal pictures

from different people (three females and three males) were used as stimuli. Their face age belonged to a vicenarian age group. For each person four different pictures were used that represented the emotional states (anger, fear, happiness, and neutral) in which the emotional state *neutral* was used as a comparison. For our study, the intact version of these faces had been manipulated using the program GIMP to add a medical mask on the mouth-nose area. The picture of the medical mask was taken from the internet and manipulated (e.g., shadows added) to create a realistic appearance. The same mask was added to each face. The different faces (masked and unmasked) including the four different emotional states were shown upright and inverted to the participants.
In sum, we obtained 2 (face sex) × 3 (individuals) × 4 (emotions) × 2 (no face mask vs. face mask) × 2 (upright vs. inverted) = 96 face stimuli.

Recording and analysis of psychophysiological data

Electroencephalogram (EEG) was recorded from 64 Ag/AgCl active scalp electrodes (10-20 arrangement) with actiCAP and BrainAmp amplifiers (Brain Products, Gilching, Germany), digitized at 1000 Hz. Horizontal electrooculogram (EOG) was recorded from two Ag/AgCl electrodes placed at the lateral canthi of both eyes. Electrocardiogram (ECG) and electromyogram (EMG) from the M. corrugator supercilii were recorded, as well, but results will be reported elsewhere. Eye movements were tracked with an SMI RED 500 eye tracking device (SensoMotoric Instruments, Teltow, Germany).

Offline analysis was performed with BrainVision Analyzer 2.2 (Brain Products, Gilching, Germany). EEG data were filtered offline with a bandpass-width of 0.1 to 25 Hz (half-power cutoff; roll-off 24 dB/Octave). The sampling rate was reduced to 250 Hz and the data were semi-automatically inspected for artefacts, excepting eye movements and blinks. Ocular artefacts were removed with independent component analysis (ICA) in semi-automatic mode. Afterwards, faulty channels were interpolated (spheric splines) and an average reference was applied. Data were then segmented from -500 to 2500 ms relative to stimulus onset, baseline corrected (-200 to 0 ms), and semi-automatic artefact rejection was performed. Then, shorter segments from -200 to 1000 ms were created and averaged per stimulus category. The N170 was quantified as mean amplitude from 128 to 188 ms after stimulus onset. The time window was determined with a collapsed localizer approach, i.e., by visual inspection of the grand average waveform, collapsed across all stimulus categories. Mean amplitudes were extracted from an average of the channels P7, P8, PO7, PO8, PO9, PO10 (Rossion & Jacques, 2008).

Procedure

The participants were first asked several questions about their current state based on a state questionnaire. Then, the participants filled in several trait questionnaires and electrodes were applied. Each participant had to perform four different tasks during the study. A resting state measurement included 5 minutes of rest with open eyes, as well as four 1-minute resting periods with eyes open-closed-open-closed-open (counterbalanced).
The first task required passive viewing of different pictures of six human faces, displaying different emotional expressions which appeared for 1000 ms on the monitor, after looking at a fixation cross, which appeared for 1250 ms (random 1000-1500 ms). A fixation of at least 500 ms on the fixation cross was required for the picture to appear, verified by the eye tracker. Each photo was repeated five times, resulting in a total of 30 repetitions per category. During the observation, the participant's brain activity was recorded. The second task represented the six faces once again but without repetitions. The participants had to look at a fixation cross, like in the first task, which attend about 1250 ms followed by the stimuli presented for 4000 ms. Then, they had to decide which emotion the face had displayed, by choosing one of four options (anger, fear, happiness, neutral). Eye movements were recorded concurrently. The third task included once again the stimuli from the MPI FACES

database (Ebner, N.C., Riediger, M., & Lindenberger U., 2010), where the several emotional expressions were shown, and it was required that the participants rate their affective experience (valence and arousal) when observing these facial expressions. The fourth and last task consisted of the heartbeat counting which is not reported in the current manuscript. The entire procedure lasted approximately 3-3,5 hours.

Statistical analysis

The first hypothesis „Configural face processing becomes impaired due to the use of medical face masks" and the second hypothesis "Emotional processing is impeded looking at faces wearing a medical face mask" consist of a dependent variable N170 and the three independent variables regarding two levels of mask *wearing a mask/no mask,* two levels of orientation *inverted/upright* and four levels of emotion *anger, fear, happiness* and *neutral*. To illustrate our results, the repeated measures analysis of Variance (ANOVA) with an alpha level .05 was used. Where Mauchly' s Test of Sphericity was significant, the Greenhouse-Geisser correction for degrees of freedom was applied.
The third hypothesis "Recognition of emotional expressions from faces wearing medical masks is impaired" consists of a dependent variable *emotion recognition* and the three independent variables regarding two levels of mask *wearing a mask/ no mask,* two levels of orientation *inverted/upright* and the four levels of emotion *anger, fear, happiness* and *neutral.* For a second time an ANOVA analysis as well as a correction for the Mauchly's Test of Sphericity was used and the Greenhouse-Geisser estimates of sphericity was applied.

Results

Hypotheses 1 and 2

For our first and second hypothesis „Configural face processing becomes impaired due to the use of medical face masks." and "Emotional processing is impeded looking at faces wearing a medical face mask." our dependent variable is the N170, which gives us indication for face and emotional processing. Here it is important to indicate that a more negative amplitude represents a larger N170, and a less negative amplitude represents a smaller N170.

The main effect of the specific emotions was significant, $F(2.28, 66.10) = 5.45$, $p = .005$, $\eta_p^2 = .16$. Post-hoc analyses were conducted using six paired *t*-tests with a Bonferroni adjusted alpha level of .0083 per test (.05/6). The results indicate that the amplitude of N170 did not differ significantly between anger ($M = -1.58$, $SD = 2.19$) and fear ($M = -1.86$, $SD = 2.35$), $t(29) = 1.78$, $p = .085$. When comparing anger ($M = -1.58$, $SD = 2.19$) and happiness ($M = -1.47$, $SD = 2.33$), there is no significant difference in the amplitude of N170, $t(29) = -.73$, $p = .47$. In addition there was also no significant difference in N170 when comparing happy faces ($M = -1.47$, $SD = 2.33$) with neutral faces ($M = -1.13$, $SD = 2.53$), $t(29) = -1.91$, $p = .07$. Moreover, the amplitude of N170 did not differ significantly from angry faces ($M = -1.58$, $SD = 2.19$) and neutral faces ($M = -1.13$, $SD = 2.53$), $t(29) = -2.30$, $p = .03$. Further, there was no significant difference between the amplitude of the N170 for fearful faces ($M = -1.86$, $SD = 2.35$) and for happy faces ($M = -1.47$, $SD = 2.33$), $t(29) = -2.46$, $p = .02$. Lastly, the amplitude of N170 was significantly smaller for neutral faces ($M = -1.13$, $SD = 2.53$) than for fearful faces ($M = -1.86$, $SD = 2.35$), $t(29) = -3.03$, $p = .005$. (see 1.1.).

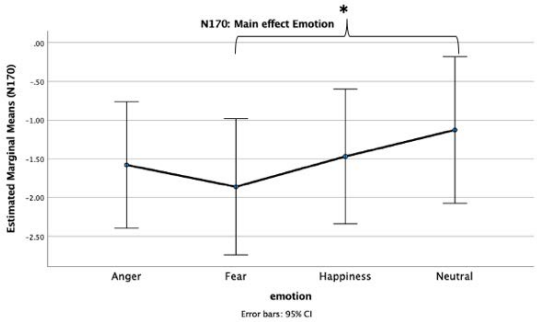


Fig. 1.1.: Graphical schematic summary showing the emotional processing. Hereby the ANOVA effect is significant ($p < .05$).
*$p < .05$

The main effect of faces wearing a mask/no mask was also significant, $F(1, 29) = 8.81$, $p = .006$, $\eta_p^2 = .23$. Faces wearing a mask (M = -

1.79, *SD* = 2.54) provoked a significantly larger N170 than faces wearing no mask (*M* = -1.23, *SD* = 2.10) (see, fig. 1.2.). The main effect of the face orientation was significant, $F(1, 29) = 53.25$, $p < .001$, $\eta_p^2 = .65$. Concerning the face orientation, the amplitude of N170 was significantly more negative for inverted faces (*M* = -2.53, *SD* = 2.58) than for upright faces (*M* = -.49, *SD* = 2.02) (fig. 1.3.).

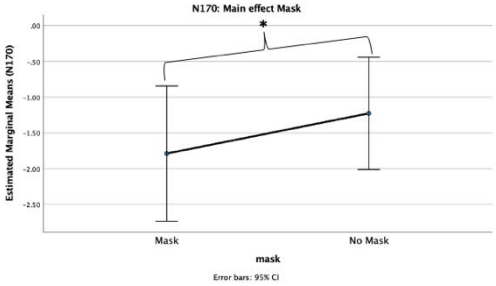


Fig. 1.2.: Graphical schematic summary showing the impact of faces wearing a mask/no mask on the facial processing. Hereby the ANOVA effect is significant ($p < .05$).
*$p < .05$

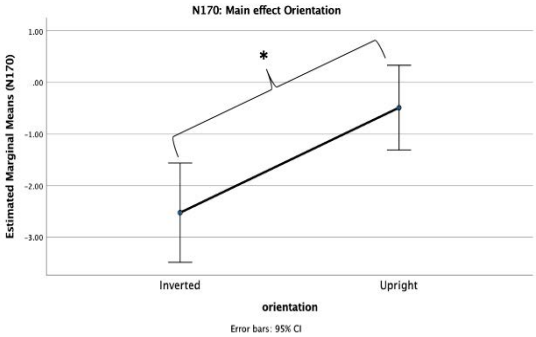


Fig. 1.3.: Graphical schematic summary showing the impact of the face orientation on the facial processing. Hereby the ANOVA effect is significant ($p < .05$).
*$p < .05$

There was a significant interaction effect between the face orientation and the specific emotions, $F(2.69, 78.18) = 11.99$, $p < .001$, $\eta_p^2 = .29$. Post-hoc analyses were conducted using four paired *t*-tests with a Bonferroni adjusted alpha level of .0125 per test (.05/4). Results suggest concerning anger, that the amplitude of N170 was significantly smaller for upright faces (*M* = .86, *SD* = 2.27) than for inverted faces (*M* = -2.30, *SD* = 2.56), $t(29) = 3.81$, $p = .001$. When looking at fearful faces, the amplitude of N170 was significantly larger for inverted faces (*M* = -2.62, *SD* = 2.75) than for upright faces, (*M* = -1.10, *SD* = 2.37), $t(29) = 4.07$, $p < 001$. Furthermore, N170 was significantly larger for inverted faces (*M* = -2.71, *SD* = 2.54) than for upright faces (*M* = -.23, *SD* = 2.32) when the specific emotion was happiness, $t(29) = 9.84$, $p < .001$. Lastly, when the specific emotion was neutral, N170 was significantly larger in inverted position (*M* = -2.48, *SD* = 2.88) than in upright position (*M* = .23, *SD* = 2.37), $t(29) = 10.13$, $p < .001$ (figure 1.4.).

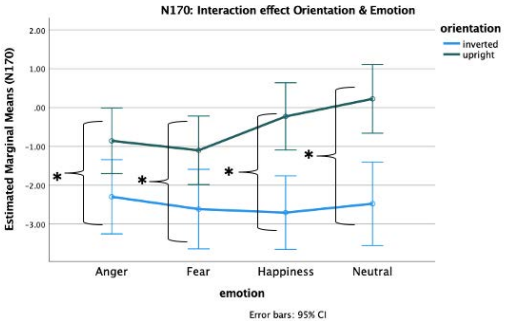


Fig. 1.4: Graphical schematic summary showing the impact of the interaction between the face orientation and the specific emotions on the facial processing. Hereby the ANOVA effect is significant ($p < .05$).
*$p < .05$

For the interaction effect between the face orientation and faces wearing a mask/no mask, the effect was also significant, $F(1, 29) = 40.18$, $p < .001$, $\eta_p^2 = .58$. Post-hoc analyses were conducted using three paired *t*-tests with a Bonferroni adjusted alpha level of .017 per test (.05/3). For faces wearing no mask results indicate that N170 was significantly larger with inverted faces (*M* = -2.72, *SD* = 2.48) than with upright faces (*M* = .27, *SD* = 2.23), $t(29) = 7.67$, $p < .001$. For faces wearing masks, the negative amplitude of N170 was significantly larger for inverted faces (*M* = -2.33, *SD* = 2.74) than for upright faces (*M* = -1.25, *SD* = 2.47), $t(29) = 4.87$, $p < .001$ (fig. 1.5.). The following paired *t*-test was conducted to find out the difference in means of the amplitude of N170 between the upright and inverted faces wearing a mask and the upright and inverted faces wearing no mask. These results show that the difference of the amplitude of N170 between inverted and

upright faces wearing no mask (M = 2.99, SD = 2,14) was significantly larger than the difference of the amplitude of N170 between inverted and upright faces wearing a mask (M = 1.08, SD = 1.21), $t(29) = 6{,}34$, $p < .001$. The interaction effect between faces wearing a mask/no mask and the specific emotions was not significant, $F(2.59, 75.11) = 1.65$, $p = .18$, $\eta_p^2 = .05$ (fig. 1.6). For the interaction effect between the face orientation, faces wearing a mask/no mask and the specific emotions, this interaction was not significant, $F(2.56, 74.12) = 2.09$, $p = .11$, $\eta_p^2 = .07$ (Fig. 1.7.; Fig. 1.8.).

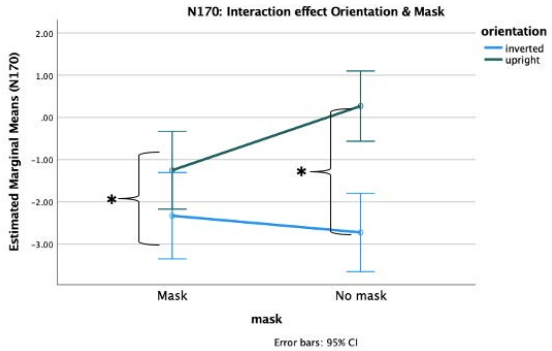


Fig. 1.5.: Graphical schematic summary showing the impact of the interaction between the face orientation and faces wearing a mask/no mask on the facial processing. Hereby the ANOVA effect is significant ($p < .05$).
*$p < .05$

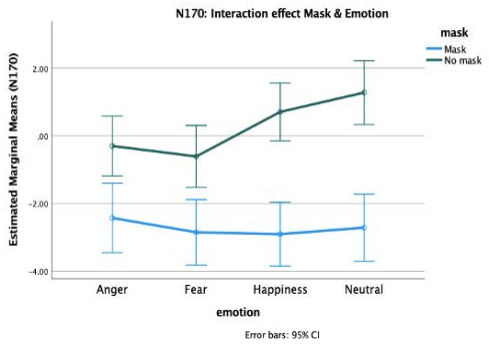


Fig. 1.6.: Graphical schematic summary showing the impact of the interaction between specific emotions and faces wearing a mask/no mask on the emotional processing.

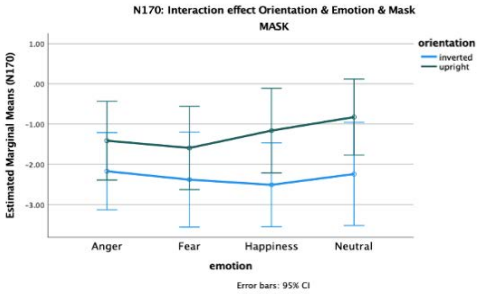


Fig. 1.7.: Graphical schematic summary showing the impact of the interaction between the face orientation, the specific emotions and faces wearing a mask on the facial processing.

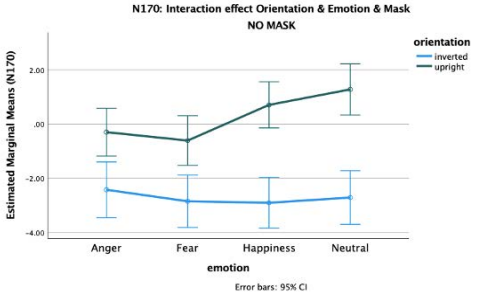


Fig. 1.8.: Graphical schematic summary showing the impact of the interaction between the face orientation, faces wearing no mask and the specific emotions.

EEG recording for anger

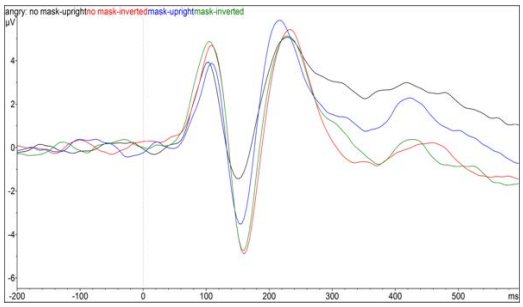


Fig. 1.9.: EEG data presenting the configural processing (N170) for angry faces for faces wearing a mask / no mask * face orientation. Hereby indicates red the EEG recording of seeing faces wearing no mask and inverted, blue stands for faces wearing mask and upright, green indicated faces wearing mask and inverted and black represents faces wearing no mask and upright. This also applies to the following graphics.

EEG recording for happiness

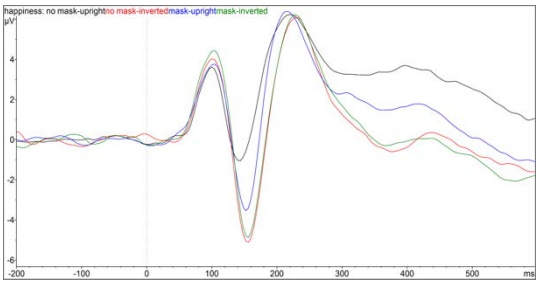


Fig. 1.10.: EEG data presenting the configural processing (N170) for happy faces for faces wearing a mask / no mask * face orientation

EEG recording for fear

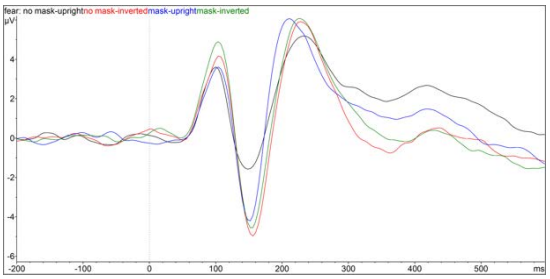


Fig. 1.11.: EEG data presenting the configural processing (N170) for fearful faces for faces wearing a mask / no mask * face orientation

EEG recording for neutral

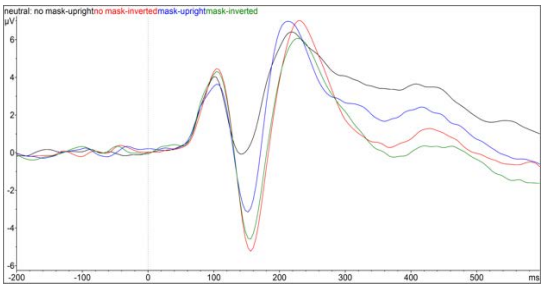


Fig. 1.12.: EEG data presenting the configural processing (N170) for neutral faces for faces wearing a mask / no mask * face orientation

Hypothesis 3

For our third hypothesis, regarding emotion recognition, our analysis showed that the main effect of the specific emotions was not significant, $F(3, 74.69) = .83$ $p = .46$, $\eta_p^2 = .027$ (Fig. 2.1.), and neither was the interaction effect between the specific emotions and faces wearing a mask/no mask, $F(3, 61.79) = 1.37$, $p = .26$, $\eta_p^2 = .04$ (Fig. 2.2.). The main effect of the face orientation was significant, $F(1, 30) = 23.254$, $p < .001$ $\eta_p^2 = .44$. Recognition rates were significantly higher for upright ($M = .99$, $SD = .02$) than for inverted faces ($M = .94$, $SD = .06$) (see, fig. 2.3.). The main effect of faces wearing a mask/no mask was also significant, $F(1, 30) = 14.831$, $p = .001$, $\eta_p^2 = .33$. Recognition rates were significantly higher for faces wearing no mask ($M = .98$, $SD = .02$) than for faces wearing a mask ($M = .95$, $SD = .06$) figure 2.4.).

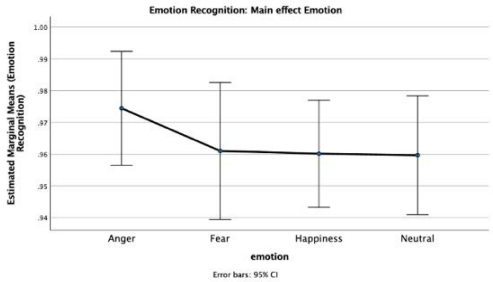


Fig. 2.1.: Graphical schematic summary showing the impact of specific emotions on the emotion recognition.

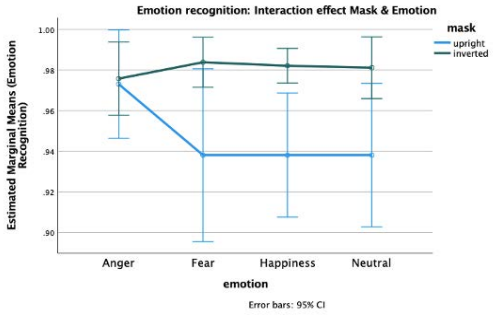


Fig. 2.2.: Graphical schematic summary showing the impact of the interaction between faces wearing a mask/no mask and the specific emotions on the emotion recognition.

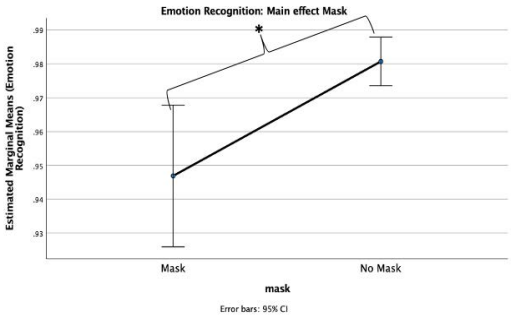


Fig. 2.3.: Graphical schematic summary showing the impact of the face orientation on the emotion

recognition. Hereby the ANOVA effect is significant ($p < .05$).
$p < .05$

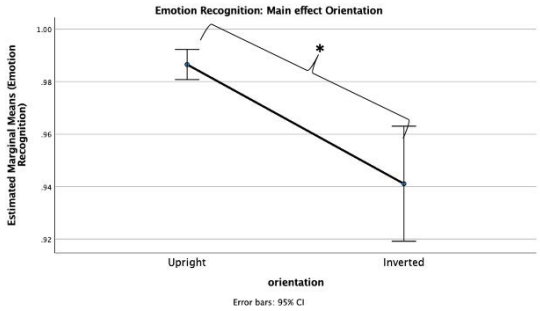


Fig. 2.4.: Graphical schematic summary showing the impact of faces wearing a mask/no mask on the emotion recognition. Hereby the ANOVA effect is significant ($p < .05$).
*$p < .05$

The interaction effect between the face orientation and faces wearing a mask/no mask, was significant, $F(1, 30) = 25.07$, $p < .001$, $\eta_p^2 = .04$. Post-hoc analyses were conducted using two paired *t*-tests with a Bonferroni adjusted alpha level of .025 per test (.05/2). Results suggest that emotional recognition of inverted faces ($M = .91$, $SD = .10$) was significantly lower than for upright faces ($M = .99$, $SD = .03$) when the faces were wearing a mask, $t(30) = 5.33$, $p < .001$. Without a mask, there was no significant difference of the emotional recognition rates between upright ($M = .99$, $SD = .02$) and inverted faces ($M = .98$, $SD = .03$), $t(30) = 1.30$, $p = .20$ (see, figure 2.5.).
The interaction effect of the specific emotions and face orientation was also significant, $F(2.47, 74.23) = 2.95$, $p = .048$, $\eta_p^2 = .09$. Post-hoc analyses were conducted using four paired *t*-tests with a Bonferroni adjusted alpha level of .0125 per test (.05/4). Results suggest that for anger, there was no significant difference between upright ($M = .98$, $SD = .04$) and inverted faces ($M = .97$, $SD = .08$), $t(29) = 1.09$, $p = .28$. Recognition of fear was significantly better for upright ($M = .98$, $SD = .04$) than inverted faces ($M = .93$, $SD = .09$), $t(29) = 2.71$, $p = .01$. Concerning the recognition of happiness, upright faces ($M = .99$, $SD = .03$) were significantly better recognized than inverted faces ($M = .93$, $SD = .08$), $t(29) = 3.72$, $p = .001$. Lastly, neutral faces were significantly better recognized in an upright position ($M = .99$, $SD = .02$) than in an inverted position ($M = .93$, $SD = .09$), $t(29) = 3.80$, $p = .001$ (fig. 2.6.).

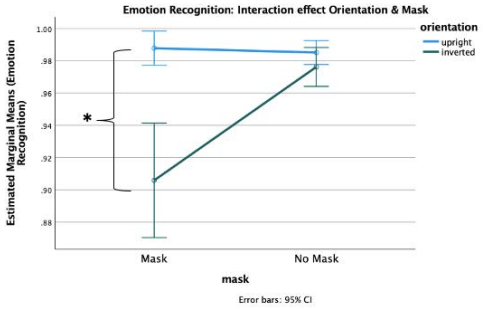


Fig. 2.5.: Graphical schematic summary showing the impact of the interaction between the face orientation and faces wearing a mask/no mask on the emotion recognition. Hereby the ANOVA effect is significant ($p < .05$).
*$p < .05$

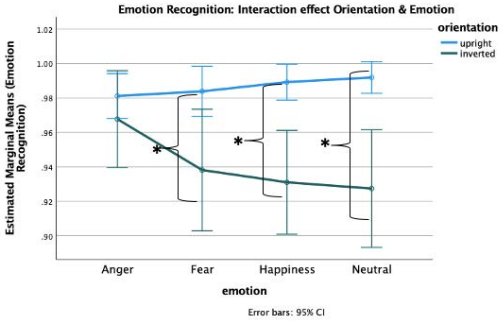


Fig. 2.6.: Graphical schematic summary showing the impact of the interaction between the face orientation and the specific emotions on the emotion recognition. Hereby the ANOVA effect is significant ($p < .05$).
*$p < .05$

If we look at the interaction effect of the 3 variables together, the specific emotions, the face orientation and faces wearing a mask/no mask, there was a significant effect for this interaction, $F(2.16, 64.86) = 3.63$, $p = .03$, $\eta_p^2 = .11$. Post-hoc analyses were conducted using eight paired *t*-tests with a Bonferroni adjusted alpha level of .00625 per test (.05/8). The following results refer to faces wearing a mask. Results suggest that for the recognition of anger, there was no significant difference between upright ($M = .98$, $SD = .06$) and inverted faces ($M = .97$,

SD = .11), *t*(30) = .63, *p* = .54. The recognition of fear was significantly better for upright (*M* = .98, *SD* = .07) than inverted faces (*M* = .89, *SD* = .18), *t*(30) = 3.59, *p* = .001. For happiness, upright faces (*M* = .99, *SD* = .04) were significantly better recognized than inverted faces (*M* = .89, *SD* = .15), *t*(30) = 4.25, *p* < 001. Lastly, neutral faces were significantly better recognized in an upright position (*M* = 1.00, *SD* = .00) than in an inverted position (*M* = .88, *SD* = .19), *t*(30) = 3.58, *p* = .001 (see, fig. 2.7.).

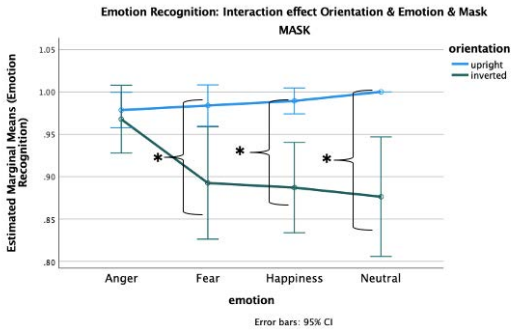


Fig. 2.7.: Graphical schematic summary showing the impact of the interaction effect between faces wearing a mask, the face orientation, and the specific emotions on the emotion recognition. Hereby the ANOVA effect is significant (*p* < .05).
**p* < .05

The following results refer to faces wearing no mask, which did not reveal any significant differences between the specific emotions and face orientation. For anger there was no significant difference between upright (*M* = .98, *SD* = .05) and inverted faces (*M* = .97, *SD* = .08), *t*(30) = 1.00, *p* = .33. Concerning the recognition of fear, inverted faces (*M* = .98, *SD* = .05) did not significantly differ from upright faces (*M* = .98, *SD* = .05), *t*(30) = .00, *p* = 1.00. Regarding happiness, there was no significant difference between upright (*M* = .99, *SD* = .04) faces and inverted faces (*M* = .98, *SD* = .03), *t*(30) = 1.49, *p* = .15. Lastly, the emotional recognition of neutral faces in upright position (*M* = .98, *SD* = .05) did not differ significantly from neutral faces in inverted position (*M* = .98, *SD* = .06), *t*(30) = .44, *p* = .66 (see, figure 2.8.).

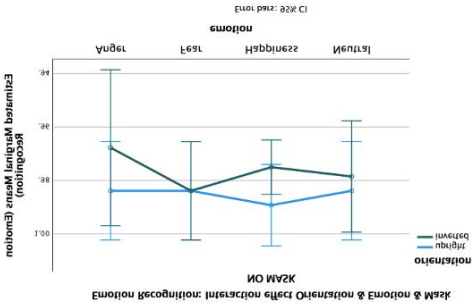


Fig. 2.8.: Graphical schematic summary showing the impact of the interaction between faces wearing no mask, the face orientation, and the specific emotions on the emotion recognition.

Discussion

Several studies have shown that emotion recognition is complicated by the masking of various facial features (Grundmann et al., 2021; Carragher, 2020; Carbon, 2020 ; Freud et al., 2021; Fischer et al., 2012). In addition, there also seem to exist different types of face processing, with configural face processing in particular gaining importance recently (Freire et al., 2000; Derntl & Carbon 2009; Schmidt et al., 2013). This was illustrated, for example, by the fact that face inversion led to difficulties in processing faces, which was shown by an increased N170. Thus, this is further evidence that the N170 is responsible for processing faces (Maurer et al., 2007; Itier & Taylor, 2004; Eimer & Holmes, 2002; Hinojosa et al., 2015). However, Hinojosa et al., (2015) also showed that there are differences in the processing of faces that depend on emotional states. The aim of our study was therefore to find out whether and to what extent the medical face masks that are part of our everyday life in the Covid 19 pandemic affect the processing of faces and emotions, as well as emotion recognition in general. To calculate our results, we used a sample size of 30 participants and employed a repeated measures analysis of variance (ANOVA).

For the first hypothesis, we found that faces in the photographs that wore a medical face mask induced a more pronounced N170 in the participants than faces that did not wear masks (mask main effect). Furthermore, we were able to demonstrate the inversion effect in the interaction effect between mask and orientation on

the N170, which was stronger for unmasked faces. Since the inversion effect can be seen as a kind of evidence for configural processing, it is clear from our results that configural processing is made more difficult by the mask, since the inversion effect was smaller in this case. Thus, we can confirm our first hypothesis. For the second hypothesis, the measurements for the main effect emotion made clear that we were also able to demonstrate different N170 outcomes depending on the emotional facial states, consistent with existing literature. Thus, the N170 was significantly stronger when looking at faces that showed the emotion fear than when a neutral condition was presented. However, we could not confirm our second hypothesis because the mask had no significant effect on the emotion processing. Thus, the inversion of faces influenced processing, but it made no significant difference to emotion processing whether the masks were worn or not. Finally, the results of our third hypothesis demonstrated that the mask was significant for emotion recognition. The mask showed a significant effect, as did the interaction between the orientation and the mask. The interaction between the emotions, the mask and the orientation were also significant, and illustrated that without the mask, all emotions were recognised equally well, regardless of whether they were presented to the participants upright or inverted. However, if the persons in the photos were wearing a mask, the recognition of the emotions became more difficult when the photos were inverted. While this was seen for the emotions of fear, happiness and neutral, no effect was found for anger. Thus, based on our results, our last hypothesis can be confirmed.

Considering previous studies as shown, we were also able to demonstrate the inversion effect and clarify that the inversion of faces leads to increased N170 amplitude and impaired processing (Yin, 1969 ; Itier & Taylor 2004; Kloth et al., 2015). Furthermore, it has been shown in previous studies that isolated eyes lead to an increased N170 (Kloth et al., 2013). We can compare this with our results, as mainly the eyes are observed when wearing masks. We could observe that upright faces in which the lower facial features such as nose and mouth were covered led to an increased N170 in the observing participants compared to when the pictured persons did not wear a mask. Of particular interest here, however, is that the amplitude was still more pronounced when looking at inverted faces. Furthermore, our results show that although the inversion effect was also observed in the mask condition, it was significantly smaller than for unmasked faces. This finding is particularly noteworthy because it illustrates that the mask and the inversion do not seem to have the same effect on the N170 of the observing participants.

Moreover, as in the meta-analysis by Hinojosa et al. (2015), we found increased N170 amplitudes for an emotional expression compared to the neutral condition. However, while anger induced the highest N170 in the meta-analysis, in our case it was the emotion fear that differed significantly from the neutral condition. We found some similarities with the existing literature in emotion recognition. For example, that emotion recognition seems to be facilitated by the eyes or the mouth, depending on the emotion (Wegrzyn et al., 2017; Blais et al., 2012; Kotsia et al., 2008). As Calbi et al. (2021) previously noted, the emotion anger seems to be recognised primarily through the eyes, which could explain our results. We found that the emotion anger was always recognised equally well, regardless of the different conditions. While Gulbetekin (2021) found in his study that the emotion fear is made more difficult by the mask, and Carbon (2020) showed that the emotion anger is more difficult to recognise by wearing a mask, we have obtained different results here as already clarified.

Limitations

As mentioned earlier, there is some concern that the mask may complicate interpersonal interactions by impairing the recognition of emotions (Carbon, 2020). However, our results indicate that clearly shown emotions seem to be well recognised on upright faces. The question here is whether difficulties can also occur when the face is seen from other angles than the front. Furthermore, it remains questionable whether emotions are shown as strongly in everyday life as they were in the photos used. It seems surprising that both the functioning of

face processing and emotion processing can be measured by the N170, but the latter does not seem to be significantly affected by the mask, while face processing is impaired. However, since emotion processing only occurs after face processing, the question arises if there might be other processes in between which compensate for the disturbed N170, so that the mask has less effect on emotion processing. Another idea would be, for example, that due to the disturbed configural processing, a different type of processing is used more. Thus, for example, featural processing could be used more intensively and therefore there might be more attention paid to details in the eye region.
Across all our results, it seems that emotions were well recognisable on upright faces despite the medical face mask, even though face processing was disturbed by the mask. As already suggested, this could be explained by the fact that we selected photos from the MPI FACES database on which the emotions were shown in an extreme form, whereas such intense facial expressions are rarely observed. This raises the question of how far the results can be transferred to the current situation. Furthermore, due to COVID restrictions, the face mask had to be edited to the face photos and were not worn while the emotions were shown. This could also affect the interpretation of the results, as it is quite possible that emotions are shown differently with a face mask on than without. In addition, we mainly studied female participants, which, regarding the existing literature, could also be a reason why the recognition of the emotion anger does not seem to have any impairments in our study. Various studies have shown that women, in comparison to men, look primarily at the eye region when processing emotions, which enables them to recognise the emotions of disgust, sadness, fear and anger particularly well (Eisenbarth & Alpers, 2011; Grundmann et al., 2021; Calbi et al., 2021; Freud et al., 2020; Carbon, 2011; Sullivan et al., 2015). In addition, regarding our sample, it can be observed that we have mainly recruited people of Luxembourgish nationality and students, which can also raise questions about the extent to which our results can be generalised to the population. In addition, we only studied one age group, which may also be a reason for the good recognition of the emotional states. The people in the photos were middle-aged, which was very close to the age of the participants and could therefore facilitate emotion recognition through the own-age effect (Susilo, 2009). Finally, participants wore masks themselves during emotion recognition, which could have led to a lack of concentration.

Further research

Based on the limitations listed, further research can be derived. For example, it would be interesting to investigate the effects of the mask on facial and emotion processing in different age groups. Here, both the age of the observers and the age of the persons shown in the photos may have an effect (Carbon, 2020; Barborik, 2012; Sullivan et al., 2015). In terms of age, it is also relevant to examine the mask effect in children. Previous research has suggested that children do not use configural processing yet, so it would be of interest to find out whether the mask affects face and emotion processing differently from adult participants (Itier & Taylor, 2004; Roberson et al., 2012; Pascalis, 2002). Moreover, expertise also appears to be important for face processing. Previous studies have shown that expertise leads to similar processing as for faces (Roberson et al., 2012; Tanaka & Curran, 2001). It would therefore be useful to conduct further studies after a few years to find out whether our processing adapts to the circumstances and the wearing of the mask. Looking at the results of Nestor et al. (2020), we seem to show less emotional expressions when we wear masks. This fact and the facial feedback hypothesis make it interesting to include another condition in our research (Sträter, 2019; Nestor et al., 2020). For example, we could investigate how participants' emotion recognition changes when they look at emotional faces with and without a mask, while they themselves wear a mask or not. Finally, it might also be worth seeing how the processing of faces and emotions differs between ethnic groups. Thus, the other-race effect shows that a different origin already seems to have an influence on processing, although it would be interesting to find out whether and to what extent

this is reinforced by a mask (Suhrke et al. 2014).

Relevance

Overall, this study has achieved research-relevant results, as no studies on the effect of medical face masks on the N170 exist to date. Moreover, it could also have increased clinical relevance: Feuerriegel et al. (2014), for example, already showed that certain groups of people with different mental disorders, e.g. Schizophrenia, seem to have different face processing. So, it seems possible that the mask could make emotion processing even more difficult.

Conclusion

Finally, our study shows that even with a medical face mask, emotions could still be recognised and processed as long as the face was presented upright and the emotional expressions shown in the photos were expressed in a very intense way. Moreover, although face processing was disrupted by the mask, there appear to be other processes between face processing and emotion recognition, so that emotion recognition may be disrupted to a lesser extent by the mask. Further studies can show to what extent the results obtained can be transferred to our everyday lives and what precisely this means for our social interactions.

References

Barborik, M. (2012). *Emotionserkennung über die Lebensspanne*. https://doi.org/10.25365/THESIS.22056

Blais, C., Roy, C., Fiset, D., Arguin, M., & Gosselin, F. (2012). The eyes are not the window to basic emotions. *Neuropsychologia*, *50*(12), 2830–2838. https://doi.org/10.1016/j.neuropdychologia.2012.08.010

Calbi, M., Langiulli, N., Ferroni, F., Montalti, M., Kolesnikov, A., Gallese, V., & Umiltà, M. A. (2021). The consequences of COVID-19 on social interactions: An online study on face covering. *Scientific Reports*, *11*(1), 2601. https://doi.org/10.1038/s41598-021-81780-w

Carbon, C.-C. (2011). The First 100 Milliseconds of a Face: On the Microgenesis of Early Face Processing. *Perceptual and Motor Skills*, *113*(3), 859–874. https://doi.org/10.2466/07.17.22.PMS.113.6.859-874

Carbon, C.-C. (2020). Wearing Face Masks Strongly Confuses Counterparts in Reading Emotions. *Frontiers in Psychology*, *11*, 566886. https://doi.org/10.3389/fpsyg.2020.566886

Carbon, C.-C., Schweinberger, S. R., Kaufmann, J. M., & Leder, H. (2005). The Thatcher illusion seen by the brain: An event-related brain potentials study. *Cognitive Brain Research*, *24*(3), 544–555. https://doi.org/10.1016/j.cogbrainres.2005.03.008

Carragher, D. J., & Hancock, P. J. B. (2020). Surgical face masks impair human face matching performance for familiar and unfamiliar faces. *Cognitive Research: Principles and Implications*, *5*(1), 59. https://doi.org/10.1186/s41235-020-00258-x

Derntl, B., Seidel, E.-M., Kainz, E., & Carbon, C.-C. (2009). Recognition of Emotional Expressions is Affected by Inversion and Presentation Time. *Perception*, *38*(12), 1849–1862. https://doi.org/10.1068/p6448

Ebner, N. C., Riediger, M., & Lindenberger, U. (2010). FACES—A database of facial expressions in young, middle-aged, and older women and men: Development and validation. Behavior Research Methods, 42(1), 351–362. https://doi.org/10.3758/brm.42.1.351

Eimer, M., & Holmes, A. (2002). An ERP study on the time course of emotional face processing: *Neuroreport*, *13*(4), 427–431. https://doi.org/10.1097/00001756-200203250-00013

Eisenbarth, H., & Alpers, G. W. (2011). Happy mouth and sad eyes: Scanning emotional facial expressions. *Emotion*, *11*(4),

860–865. https://doi.org/10.1037/a0022758

Feuerriegel, D., Churches, O., Hofmann, J., & Keage, H. A. D. (2015). The N170 and face perception in psychiatric and neurological disorders: A systematic review. *Clinical Neurophysiology*, *126*(6), 1141–1158. https://doi.org/10.1016/j.clinph.2014.09.015

Fischer, A. H., Gillebaart, M., Rotteveel, M., Becker, D., & Vliek, M. (2012). Veiled Emotions: The Effect of Covered Faces on Emotion Perception and Attitudes. *Social Psychological and Personality Science*, *3*(3), 266–273. https://doi.org/10.1177/1948550611418534

Freire, A., Lee, K., & Symons, L. A. (2000).

The Face-Inversion Effect as a Deficit in the Encoding of Configural Information: Direct Evidence. *Perception*, *29*(2), 159–170. https://doi.org/10.1068/p3012

Freud, E., Stajduhar, A., Rosenbaum, R. S., Avidan, G., & Ganel, T. (2020). The COVID-19 pandemic masks the way people perceive faces. *Scientific Reports*, *10*(1), 22344. https://doi.org/10.1038/s41598-020-78986-9

Grundmann, F., Epstude, K., & Scheibe, S. (2021). Face masks reduce emotion-recognition accuracy and perceived closeness. *PLOS ONE*, *16*(4), e0249792. https://doi.org/10.1371/journal.pone.0249792

Gulbetekin, E. (2021). *Effects of Mask Use and Race on Face Perception, Emotion Recognition, and Social Distancing During the COVID-19 Pandemic* [Preprint]. In Review. https://doi.org/10.21203/rs.3.rs-692591/v1

Hinojosa, J. A., Mercado, F., & Carretié, L. (2015). N170 sensitivity to facial expression: A meta-analysis. *Neuroscience & Biobehavioral Reviews*, *55*, 498–509. https://doi.org/10.1016/j.neubiorev.2015.06.002

Hole, G. J., George, P. A., & Dunsmore, V. (1999). Evidence for Holistic Processing of Faces Viewed as Photographic Negatives. *Perception*, *28*(3), 341–359. https://doi.org/10.1068/p2622

Itier, R. J., & Taylor, M. J. (2004). Face Recognition Memory and Configural Processing: A Developmental ERP Study using Upright, Inverted, and Contrast-Reversed Faces. *Journal of Cognitive Neuroscience*, *16*(3), 487–502. https://doi.org/10.1162/089892904322926818

Kloth, N., Itier, R. J., & Schweinberger, S. R. (2013). Combined effects of inversion and feature removal on N170 responses elicited by faces and car fronts. *Brain and Cognition*, *81*(3), 321–328. https://doi.org/10.1016/j.bandc.2013.01.002

Kotsia, I., Buciu, I., & Pitas, I. (2008). An analysis of facial expression recognition under partial facial image occlusion. *Image and Vision Computing*, *26*(7), 1052–1067. https://doi.org/10.1016/j.imavis.2007.11.004

Nestor, M. S., Fischer, D., & Arnold, D. (2020). "Masking" our emotions: Botulinum toxin, facial expression, and well-being in the age of COVID-19. *Journal of Cosmetic Dermatology*, *19*(9), 2154–2160. https://doi.org/10.1111/jocd.13569

Roberson, D., Kikutani, M., Döge, P., Whitaker, L., & Majid, A. (2012). Shades of emotion: What the addition of sunglasses or masks to faces reveals about the development of facial expression processing. *Cognition*, *125*(2), 195–206. https://doi.org/10.1016/j.cognition.2012.06.018

Sträter, L. (2019). *Erkennung und Verarbeitung emotionaler Gesichtsausdrücke bei Patienten mit akuter peripherer Fazialisparese und gesunden Kontrollprobanden: Betrachtung von Verhaltensdaten und ereigniskorrelierten Potenzialen* [Friedrich-Schiller-Universität Jena]. https://doi.org/10.22032/DBT.39804

Sullivan, S., Campbell, A., Hutton, S. B., & Ruffman, T. (2017). What's good for the goose is not good for the gander: Age and gender differences in scanning emotion faces. *The Journals of Gerontology:*

Series B, *72*(3), 441–447. https://doi.org/10.1093/geronb/gbv033

Tanaka, J. W., & Curran, T. (2001). A Neural Basis for Expert Object Recognition. *Psychological Science*, *12*(1), 43–47. https://doi.org/10.1111/1467-9280.00308

Wegrzyn, M., Vogt, M., Kireclioglu, B., Schneider, J., & Kissler, J. (2017). Mapping the emotional face. How individual face parts contribute to successful emotion recognition. *PLOS ONE*, *12*(5), e0177239. https://doi.org/10.1371/journal.pone.0177239

Suhrke, J., Freitag, C., Lamm, B., Teiser, J., Fassbender, I., Poloczek, S., Teubert, M., VÃ¶hringer, I. A., Keller, H., Knopf, M., Lohaus, A., & Schwarzer, G. (2014). The other-race effect in 3-year-old German and Cameroonian children. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00198

Susilo, T., Crookes, K., McKone, E., & Turner, H. (2009). The Composite Task Reveals Stronger Holistic Processing in Children than Adults for Child Faces. *PLoS ONE*, *4*(7), e6460. https://doi.org/10.1371/journal.pone.0006460

Yin, R. K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology*, *81*(1), 141–145. https://doi.org/10.1037/h0027474

https://www.gimp.org/

Using Robots and Tablets in Education for Children and Adolescents with Autism Spectrum Disorder: A Study Comparing Behaviour, Cognition and Preferences

Catarina Godinho Coelho, Meret E. Hoffmann, Stefan Martins, Eva A. Nittenwilm, Dana Paulus and Maria Vintila

Supervision: Doctoral Researcher Louise Charpiot

One of the key characteristics of Autism Spectrum Disorder is a deficit in social communication and interaction. So far, research has shown the use of robots to be successful in therapeutic and educational interventions. This study examines various reactions of autistic children towards robot and tablet storytellers. Our sample consisted of 11 male children with ASD (N = 11), ranging from 9 to 17 years old. Eye gaze, restricted and repetitive behaviors and proximity were examined in two conditions: a robot and a tablet telling short stories. Additionally, the subjects' attention and memory were analyzed with the help of simple questions. Finally, the participants were asked to state their preferences at the end of the experiment. We hypothesized that they would present more willingness to interact with the robot. However, no significant differences were found between the two conditions, nor between their cognitive performances. On the other hand, our results suggest a clear preference for the tablet (72.7%), as opposed to the robot (27.3%). In the future, 90.9% of the participants stated that they would rather work with a tablet, as opposed to their teacher (9.1%) or a robot (0%). This could be explained by familiarity effects, as most participants had prior experience in working with tablets, but not with robots. These findings are interesting for the future of Robot Assisted Therapy, which could potentially include tablets, making education and care more accessible to the wider ASD population. Nevertheless, longitudinal studies with larger samples are needed to support our results.

1 Introduction

1.1 Autism Spectrum Disorder

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition specified under a dyad of characteristics, namely difficulties in social communication and interaction, as well as restricted and repetitive patterns of behaviour, interests or activities (Lai et al., 2014; American Psychiatric Association, 2013). The severity and nature of symptoms can vary interindividually, but common impairments of children with ASD include difficulties in communicating one's emotions, under-standing the emotions of others, as well as a lack of eye contact and joint attention behavior (Waterhouse, 2013).

Furthermore, children with ASD generally experience difficulties when it comes to verbal and non-verbal communication, and they can have a high sensitivity to physical contact (Waterhouse, 2013; Lord et al, 2018). In addition, the restrictive and repetitive behaviors (RRB's) that people with ASD tend to exhibit can be difficult to experience for both autistic people and their respective families (Gabriels et al., 2005). When shown in an excessive manner, RRB's can inhibit the person's capacity to gain new skills (Dunlap et al., 1983), stigmatize them, and overall reduce their chances of positive interaction (Durand & Carr, 1987; Lee et al., 2007; Loftin et al., 2008). In previous studies, RRB's were defined as a wide spectrum of behaviors, ranging from self-injurious gestures, to stereotyped motor mannerisms, echolalic speech, the demand of sameness and urges to sensory interests and ab-normalities (Bodfish et al., 2000; Lewis & Bodfish, 1998; Turner, 1999).

Epidemiological studies show that in child and adolescent populations, ASD ranges at around a value of 1% (Baird et al., 2006; Baron-Cohen et al., 2009). Although there appears to be a high variability in ASD prevalence across sites, it has also been demonstrated that there is a similar distri-bution by race and ethnicity (Center for Disease Control and Prevention, 2018). In fact, studies show a mean ratio prevalence estimation of 4.2:1 comparing males and females (Fombonne, 2009). The overall substantially great prevalence of ASD in young populations increases the demand for diagnostic procedures and forms of therapy in the health sector (Baird et al., 2006).

1.2 Use of tablets and robots in ASD interventions

Besides parent-mediated interventions, a common therapeutic method for children with ASD is Applied Behaviour Analysis (ABA). ABA is a behavioural intervention which focuses on playing, social inter-action, communication initiated by the child and a natural reward system (Lord et al., 2018). Studies regarding animal-assisted intervention for ASD have also shown positive outcomes, although there are methodological weaknesses and limitations (O'Haire, 2012). Alongside these forms of intervention, there are several behavioural and social therapeutic possibilities, such as benefitting from social skills groups or parent-child interaction therapy (Lord et al., 2018).

All the aforementioned approaches have shown success. However, some forms of therapy are very time consuming and require many resources, often including different professionals in interdisciplinary teams. Furthermore, because of the high variability of severity and symptomatic demands, the need of individual therapy requires a lot of effort (Thill, 2012). This financial and temporal problem could be solved by a wide range of tasks in robotics acting as a diagnostic or behaviour eliciting agent, as a social mediator and actor, as a personal therapist or as a playmate (Cabibihan et al., 2013; Waterhouse, 2013; Vanderborght, 2012).

In the context of ASD interventions, research on Social Assistive Robots (SAR) has shown promising results, focusing on the intersection between assistive robotics (providing physical assistance) and socially interactive robotics (social and nonphysical interaction). Several studies have in fact researched the use of SAR in education or therapy (Thill et al., 2012). The use of this technology has been shown to improve engagement in social interaction, including imitation, eye contact, turn-taking and self-initiation (Feil-Seifer & Matarić, 2011; Scassellati et al., 2012; Cabibihan et al., 2013). As the end users of SAR are indivi-duals with disabilities, research in this field focuses on their use in common environments, such as schools or hospitals (Feil-Seifer & Matarić, 2011).

An advantage of robot-assisted therapy is the combination of human-like-cues and an object-like simplicity. This makes the interaction with autistic children more attractive, as they generally prefer predictability and show preference in interacting with objects. In this way, it is possible to avoid high-complexity situations which could distress the child (Srinivasan et al., 2016; Melo et al., 2019). In fact, Robins et al. (2006) found that the physical proximity of children with ASD increased when they thought they were interacting with a robot instead of a human. This has been shown to mean attachment and is therefore associated with a positive feeling (Dissanayake & Crossley, 1996).

Moreover, it has been shown that the eye gaze of children with ASD increases when dealing with a robot (with human features), in comparison to humans (Robins et al., 2006). Eye gaze is the nonverbal phenome-non which is most present in human interaction (Argyle, 1972). Similar to gestures, eye gaze can communicate a person's mental state, increase verbal communication and place emphasis on what has been said in the conversation (Goldin-Meadow, 1999). In the context of ASD, eye gaze can also be interpreted as a desire to communicate (Ho et al., 2015). However, novelty effects could possibly have interfered with these findings. As a matter of fact, a study by Michealis et al. (2021) suggests that every interaction between a human and a robot is in some way influenced by novelty effects, leading to an increase of

interaction and interest when new technology is introduced. Another study found that this effect subsided after repeated interaction with the robot in question (Saad et al., 2020).

As acknowledged by previous research, children with ASD tend to prefer to interact with robots rather than with humans (Costa et al., 2018). This phenomenon could be due to the intricacy and quantity of facial expressions as well as body language emitted by humans. It has in fact been shown that autistic children generally feel more at ease when interacting with a robot which only has a few simple humanoid characteristics (LuxAI, 2019). In addition, findings demonstrate that both self- and parent reports of RRB's show a significant positive relationship between RRB's and anxiety (Baribeau et al., 2019; Joyce et al., 2017).

Research has also shown that working with a robot can ameliorate performances in cognitive processes such as memory and verbal fluency. This study was conducted with participants who had mild cognitive impairment (Pino et al., 2020), which could also be relevant for participants with ASD, whose cognitive impairment has been highlighted earlier in the introduction.

As opposed to the use of robots in the context of ASD intervention, few studies can be found concerning the use of tablets. Nevertheless, the use of mobile digital technology is expanding, aiming more specifically to improve healthcare and psychotherapy. This method offers opportunities for the daily use in educational contexts for children with ASD (Mechling, 2007). Overall, students with ASD perceive the use of mobile technology as useful, fun and helpful. It has also been shown to improve socio-cognitive functioning (Esco-bedo et al., 2012; Fage et al., 2018).

Besides that it is important to note that children with ASD have been shown to perform better when interacting with robots rather than with tablets, particularly in cognitive aspects such as comprehension of social stories (Pop et al., 2013; Louie et al., 2020). As previously highlighted, the use of robots has indeed been associated with higher cognitive performances, which could ex-plain this difference.

1.3 Aims and hypotheses

Using robots and/or tablets could be an efficient manner to provide better care and education for autistic children. However, as the use of these two devices differs greatly in terms of cost and time (research, training, etc.), it is reasonable to clarify the differences between the two media in their use with ASD. We have thus decided to observe how autistic children differ in their behavioral, cognitive and evaluative reactions towards a robot telling a story, in comparison to a tablet telling a similar story. Based on previous research, our hypotheses are the following:

H1. Children with ASD are more likely to approach the storyteller if the latter is a robot, in comparison to a tablet (physical proximity).

H2. Children with ASD are more likely to present a longer eye gaze if the storyteller is a robot, in comparison to a tablet.

H3. Children with ASD are less likely to present restricted and repetitive behaviors during the interaction with the tablet in comparison to the robot.

H4. Children with ASD are more likely to remember details of the story if the storyteller is a robot, in comparison to a tablet (attention/ memory).

H5a. Children with ASD are more likely to prefer the story told by the robot rather than the story told by the tablet.

H5b. In the future, children with ASD would prefer to work with a robot rather than with their main teacher or with a tablet.

2 Methods

2.1 Subjects

Our sample consisted of 11 children and adolescents with Autism Spectrum Disorder, ranging from 9 to 17 years old (M= 12.45, SD = 3.30). The recruitment took place in a specialized center, focusing on the diagnosis and care of autistic youth. In fact, the participants were part of four separate classes (representing two

different grades) which had two teachers each. Key exclusion criteria were a typical neurodevelopment, not understanding Luxemburgish or French, as well as not being present on that school day.

In order to authorize the children to participate in the study, all parents were required to sign a detailed consent form; the latter was also signed by the participants when possible. Although 13 consent forms were completed in total, two sessions could not be finished as the subjects showed signs of distress. One of them refused to start the experiment, whilst the other one ran out of the room after a few seconds of interaction with the robot. Thus, a total of 11 participants (all male) comple-ted the study. A demographic questionnaire (detailed in 2.3.1) also allowed us to gather the following data :

The majority of our participants were Luxembourgish (54.5%), whilst 36.4% were Portuguese. However, most subjects reported Portuguese to be their mother tongue (54.5%). The second most reported native language was Luxemburgish, spoken by 18.2% of the children. Another 18.2% reported having a mother tongue other than Luxemburgish, Portuguese, French or German. Despite these differen-ces, all participants were able to under-stand Luxemburgish and/or French. In fact, 72.7% of the participants chose Luxem-bourgish as their language of experi-mentation. The other 27.3% chose French.

Furthermore, all subjects had formerly been diagnosed with Autism Spectrum Disorder by a health professional. In addition, an SRS questionnaire was completed by the children's parents in order to assess their level of social responsiveness (more details in 2.3.1). The mean T-score of the sample was 77.55 (*Min*= 61; *Max*= 94; *SD*= 10.77), and the majority of scores were associated with severe deficits in social interaction (54.5%). Moreover, 1 out of 11 children (9.1%) had another disability, notably a Specific Language Impairment. The rest of the subjects (90.9%) were not reported to have any comorbidities. Most parents also indicated not to know the IQ level of their children (90.9%), with the exception of one participant whose IQ was reported to be normal (9.1%). Finally, 82.8% of the subjects were considered moderately verbal by their parents.

On the other hand, our demographic questionnaire included a question con-cerning any potential specific interests. Three children had none (27.3%), whilst 8 (72.7%) had at least one. The latter included music, films, cooking, electronics, games, Google Maps, informatics, mathe-matics, and the use of a tablet. Table 1 summarizes the participants' main charac-teristics, notably their age, verbality level, SRS severity range, as well as their past experiences with a robot and/or a tablet. This last factor was included in order to examine any potential familiarity effects during their interactions with both story-teller-devices.

Table 1: Distribution of socio-demographic characteristics

Age	Verbality[a]	SRS2-Severity	Experience Tablet & Robot
13	MV	Mild	Both
17	MV	Severe	None
17	V	Moderate	Tablet
14	MV	Moderate	Tablet
9	MV	Severe	Tablet
10	MV	Severe	Both
9	MV	Moderate	Tablet
11	MV	Moderate	Tablet
10	V	Severe	Tablet
10	MV	Severe	Tablet
17	MV	Severe	Tablet

[a]MV = Moderately Verbal; V = Verbal

Finally, the questionnaires indicated that at the time of the experiment, 9.1% of the participants had been working with their current educator for less than 6 months, and another 9.1% for one to two years. Meanwhile, the majority of the children (63.6%) had been working with their teacher for more than two years. This factor was also included in order to assess potential familiarity effects, knowing that the educator would be present during the experiment.

2.2 Research Design and Procedure

We chose to conduct a field experiment, which took place in a specialized center focusing on the care of children with ASD. A separate classroom was used for the adolescents, whilst the younger children were placed in an adjoining room. Each participant took part in two trials (separate interactions with a robot and a tablet), which were randomized. Alternately, a subject was exposed to the tablet and then to the robot, whilst the next participant interacted with the robot and then with the tablet. No control group was included. Altogether, the experiment lasted for approximately ten minutes, and took place over the course of 4 days. Before starting the experiment, two team members prepa-red the room by setting up two cameras at different angles and building a square zone on the floor with white tape. In this way, the participants could stay in sight of the cameras at all time (see Figure 1). The experimenters also placed chairs for themselves and the educator in the corner of the room, behind the participant. No sitting options were provided for the children in order to encourage them to keep moving and possibly interact with the devices.

Each subject was accompanied into the room by one of their main teachers, where they met the two experimenters. It was then explained to the children that a study would be conducted and that they would have to listen to a short story. They were also advised to remain within the taped square. After the brief instructions were understood, the participants were told that the cameras had to be readjusted, when in reality their behaviour was being observed for 90 seconds. In the cases where 90 seconds were deemed too long, the study conductors talked to the children and explained the procedure in more detail. One of the experimenters then started playing the script of the robot or the tablet, in accordance with the randomization.

Figure 1: *Experimental setup*

During the experiment, the study conduc-tors and the teacher stayed in the back of the room, not interacting with the child. Based on the previously collected demo-graphic data, the script was played in the preferred language of each child (either Luxembourgish or French). The stories had been pre-recorded by a research team member who was not present during the experiment. Furthermore, the robot ("Zenbo Junior II" by Asus) was controlled via an app ("EMMA"). In this way, the script could be played in response to each child's verbal answers. The tablet (Microsoft Surface) had also been pre-programmed with an application named "OpenSesame". In both cases, the script started with simple questions in order to engage with the children.

Two stories had been prepared: one for each device. They were similar in emotional content and structure but did not include the same characters or main events. During the stories, behavioral data was collected once again. The narrative content was coupled with facial expressions for both the robot and the tablet, so that matching facial expressions would appear on the screen. Due to copyright reasons, we did not use the original "Zenbo Junior II" face animation, but a cartoon-like face anima-tion. Both stories touched on the topic of the forest and animals. The one told by the robot described an outdoor race between a rabbit and a turtle, which was surprisingly won by the latter. The story told by the tablet described a wolf, who chose to eat a piece of meat although it belonged to a fox.

After the story, the device asked the child three simple questions about its content (i.e. "Who

were the main characters? ") in order to measure their attention and memory. The children could either answer by talking to the device or by pointing to pictures, sometimes touching them as well. Then, the second trial began, with the same procedure as the first trial: 90 seconds of free interaction with the device, followed by the story and the interview focusing on attention and memory. After both trials were over, the children were asked three questions in order to evaluate their experience and to state their preferences (specified in 2.3.2). Depending on each evaluative question, the participants could point their finger towards the devices, the provided pictures or their educator. They could also respond verbally.

2.3 Measures

The study was based on socio-demographic, cognitive, evaluative and behavioral data. Firstly, the participants' parents were asked to complete a demographic questionnaire, along with the Social Responsiveness Scale 2 (SRS-2). Evaluative and cognitive data (focusing on attention and memory) was also gathered in the form of interview-type questions. Finally, observational data was collected in the form of video material. More specifi-cally, behaviors such as eye gaze, proxy-mity, and restricted and repetitive behaviors were measured.

2.3.1 Individual characteristics

We constructed a paper questionnaire in French (translated to English for one of the participants), which was filled out by the subjects' parents. They were firstly asked to indicate basic demographic information, i.e. the name, date of birth, gender, nationality and mother tongue of their child. The questionnaire also inquired about the participant's ASD diagnosis and severity level, as well as the presence of any other disabilities. Furthermore, the parents were asked about the children's IQ-level, along with their verbal ability and their specific interests, when existing. Lastly, the questionnaire included questions regarding the participants' past experience with robots, tablets, and with their current educator.

Moreover, the subjects' guardians were asked to complete the SRS-2 question-naire, which is the second edition of the Social Responsiveness Scale. It is a commonly used screener for ASD. The SRS-2 was mainly handed out in French, with the exception of one participant, whose parents preferred the English version. The SRS indirectly measures characteristics related to ASD as perceived by third party observers, in this case the parents of the participants. The questionnaire consists of 65 items divided into five treatment subscales and provides an overall total score. For the purpose of our study, only the total score was considered relevant. The items of the SRS-2 were evaluated on a four-point Likert-type scale ranging from *not true* (1), *sometimes true* (2), *often true* (3) to *almost always true* (4). During the assessment of the questionnaires, four different results were possible, each suggesting a severity range and its clinical significance. A T-score of 76 or higher suggests severe clinical significance regarding social deficits, whilst a T-score ranging from 66 to 75 is considered as moderate, and a T-score from 60 to 65 is considered mild. Finally, a T-score of 59 and below indicates that the individual has no social difficulties indicative of a possible ASD diagnosis. The SRS-2 is a reliable and valid scale; internal consistency across items is in fact at a good level (Constantino & Gruber, 2012).

2.3.2 Interview data

Cognitive data was collected as the devices asked questions regarding the content of each story. This was done in order to measure attention and memory. The three questions were simple and touched upon main characters and events. Before the experiment, an educator who is usually in charge of autistic children had validated the content and the difficulty level of both stories and all questions. In order to simplify the questions even further, our team members provided pictures for some questions, depicting every possible answer. In most cases, two options were available, with the exception of one question with four options. The robot asked the children about the main character and event of its story (rabbit and turtle; outdoor race), as well as the winner (rabbit). The

tablet's questions were very similar but adapted to its different plot (wolf and fox).

After both trials, three evaluative questions were asked. In order to operationalize evaluative variables, a similar study chose to use simple questions (Kose-Bagci et al., 2009). The main difference between this study and ours was the use of a Likert-Scale. In our case, such a scale was deemed inappropriate by our subjects' teachers, who believed the children would have had difficulties expressing their opinion in this manner. We thus constructed questions that were more adapted to the participants' level of compre-hension and cognition, following their educators' advice. Firstly, the experiment-ters asked which story the children preferred; they could answer by pointing towards the device. Secondly, the partici-pants were asked whether they found the stories easy or not, with the help of two pictures depicting a "thumbs up" and a "thumbs down". Lastly, the team members asked the subjects if they would prefer to work with a robot, a tablet or their main teacher in the future. They could answer by pointing towards their preference.

2.3.3 Observational data

In order to collect observational data, we recorded every session using two cameras, placed at different angles. After the experiment, a team member performed a second-by-second analysis of the videos. Three behavioral criteria were measured, namely proximity to the device, eye gaze and restricted and repetitive behaviors (RRB's). In the beginning, the children had 1:30 minutes of "silent time" in order to get familiar with the material and interact with it freely. Each child's level of proximity was assessed during the initial silent time, so that the results wouldn't be biased by the tactile modality of the tablet. In this way, their approach and retreat behavior was measured with the help of a four-point Likert-type scale ranging from *very far* (4), *far* (3), *close* (2) to *very close* (1). Eye gaze frequency and duration was also measured during the silent part of the experiment and the story. In this case, total frequency refers to how often the children looked at the device, whilst total duration refers to how long they looked at the device. Finally, the frequency of each participant's RRB's was collected during the silent and storytelling parts of the study. Here, RRB's were defined as any purposeless behavior exhibited at least 3 times in a row by the children. In our study, examples of such chains of behavior included repeating seemingly meaningless words or sentences, pointing at no particular object or person, and more. The behaviors were later coded by number of chains and number of behaviors in one chain (Costa et al., 2018).

3 Analysis

The behavioral data was coded by one team member, who performed a second-by-second analysis of each experiment. In order to assess the children's proximity to the storyteller, a 4-point Likert-scale was used (1 = *very close*; 2 = *close*; 3 = *far*; 4 = *very far*). For further analysis, we selected the point of the scale in which the children spent the most time (mode). We also coded the subjects' number of eye gazes towards the storyteller, as well as the duration of each gaze. We divided the frequency by the total duration of eye gazes so that we could include both measurements in a single value.

Furthermore, in order to analyze the children's RRB's during each interaction, our team member coded the number of RRB chains and number of repetitions per chain. The averages of these values were used for the statistical analysis. In other to assess the children's attention and memory, we coded their answers by assigning 1 point for every correct answer, and no point (0) for every wrong answer. We used the sum of right answers to obtain a personal score for further analysis. When coding the first evaluative question, we assigned numbers to each answer option. In fact, we coded the answer "robot" using number 1, and the answer "tablet" using number 2. For the second question, we assigned number 1 to the answer "yes", whilst number 0 was assigned to the answer "no". Likewise, the last evaluative question was coded by assigning numbers to each answer option. We coded the answer "robot" using

number 1, the answer “tablet” using number 2 and the answer “teacher” using number 3.

According to the central limit theorem of Lindenberg and Lévy (Field, 2009) our sample size (*N=11*) does not allow to assume normal distribution (*N < 30*). Therefore, we only used non-parametric tests for our statistical analysis.

4 Results

In order to statistically analyze our behavioral data, non-parametric Wilcoxon signed-rank tests were used.

H1. When it comes to proximity, our results show that children with ASD tend to approach the robot (*M= 2.73; SD= 0.77*) more than they tend to approach the tablet (*M= 2.82; SD= 0.98*). However, this difference is not statistically significant, *Z= -0.26; p= 0.49; r= 0.1* (see Figure 2).

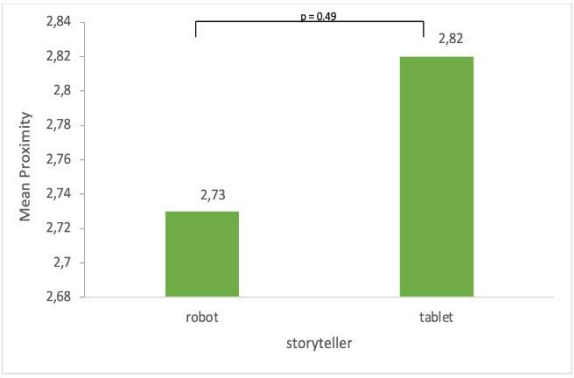


Figure 2: *Relationship between proximity and storytelling conditions*

*N.B. The lower the score, the lower the proximity between the participant and the storyteller (p=0.*49*).*

H2. Regarding the childrens’ eye-gaze towards the tablet and the robot, our results suggest that the participants looked at the tablet more often and for longer periods of time *(M*= 97.09*; SD*= 25.84*)* than at the robot *(M*= 90.41*; SD*= 47.02*)*. However, this difference is not statistically significant, *Z*= -0.45*; p*=0.35*; r*= 0.13 (see Figure 3).

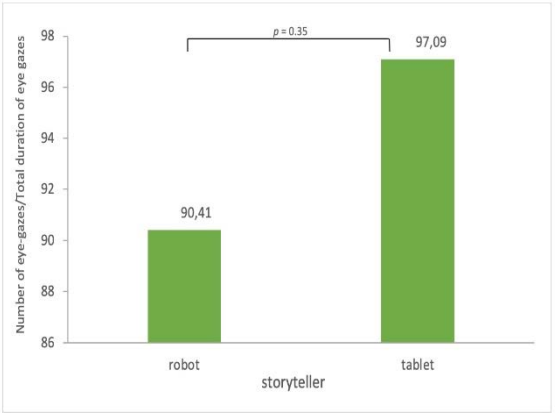


Figure 3: *Relationship between eye-gaze and storytelling conditions (p= 0.35)*

H3. When analyzing the frequency of the children’s RRB’s, a significant difference was found between both conditions,*Z*= -2.55*; p*= 0.004*; r*= 0.76 (see Figure 4) . In fact, the children showed more RRB’s during the interaction with the robot *(M*= 6.16*; SD*= 4.67*)* than during the interaction with the tablet *(M*= 3.54*; SD*= 3.62*)*.

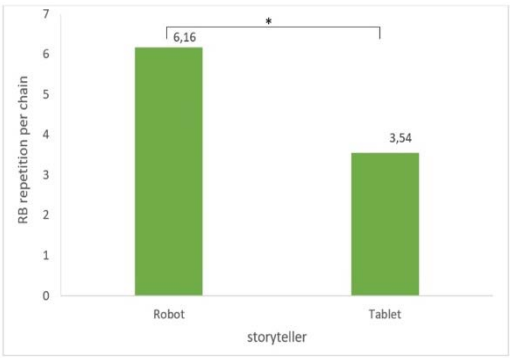


Figure 4: *Relationship between restricted and repetitive behaviors and storytelling conditions (p<0.05)*

H4. Regarding the cognitive data, our results show that the questions asked by the tablet (*M*= 1.64; *SD*= 0.67; *M*in= 1; *Max*= *3*) were more often answered correctly than the questions asked by the robot (*M*= 1.27; *SD*= 1.01; *Min*= 0; *Max*= 3).

H5a. When it comes to the evaluative data, we found that more children preferred the story told

by the tablet compared to the story told by the robot (see Figure 5).

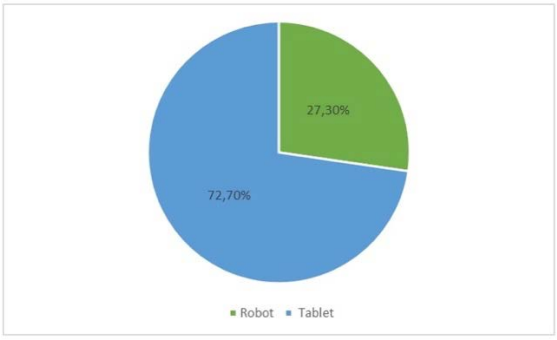


Figure 5: *Percentage of the subjects' story preferences*

H5b. Finally, our results show that in the future, all participants but one (90.9%) would rather work with a tablet than with their teacher or a robot. A single subject declared that he would prefer to work with his teacher (9.1%), whilst none of the children chose the robot as their future preference (see Figure 6).

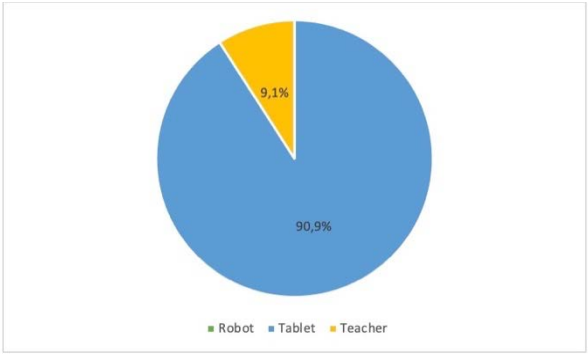


Figure 6: *Percentage of subjects' preferred device for the future*

4.1 Further Analysis

In addition to our hypotheses, we searched for a potential correlation between the participants' age and their preferred story. Spearman's rho was thus used. Although the results show that there is indeed a positive correlation between the two variables (ρ= 0.17), the correlation is weak and not statistically significant *(*p= 0.31). Furthermore, we searched for potential correlations between the participants' SRS severity range and their cognitive scores, as well as their preferred story. A positive correlation was found between the subjects' cognitive scores and their SRS severity range (ρ= 0.17*),* but this correlation was not statistically significant (p= 0.31). Likewise, another positive correlation was found between the children's preferred story and their SRS severity range (ρ= 0.18*)*, but again, it was not statistically significant (p= 0.30). It is also important to note that 10 out of 11 participants (90.9%) rated the stories as easy. Only one participant (9.1%) rated the stories as difficult.

5 Discussion & conclusion

The aim of this study was to compare the cognitive and behavioral attitudes of children with ASD towards robot and tablet storytellers. The children's eye-gazes, their proximity to the devices as well as their amount of restrictive and repetitive behaviors were analyzed and then compared. In addition to the behavioral data, the subjects' attention was also measured through their capacity to recall simple elements of each story. The children's personal evaluation of the interaction was taken into account as well. The findings of the study and their implications will now be discussed.

The present study suggests that children with ASD are as likely to approach a tablet as they are to approach a robot. Indeed, no statistically significant difference was found between the subjects' proximity to each device. Likewise, the analysis showed no significant difference between the subjects' eye gazes in each condition, suggesting that autistic children are as likely to establish eye contact with a robot as with a tablet. It is important to note that the visual stimuli were the same for both devices (Emma face animation), which may have contributed to these similar results. Our two initial hypotheses regarding proximity and eye gaze have thus been rejected. These findings are particularly interesting as they demonstrate that the subjects of this study were willing to

communicate almost equally with both devices. In a therapeutic setting, this might indicate that the use of tablets and robots could improve social interaction at a similar rate, as long as the same scripts and animations are used. In fact, if children with ASD are more likely to engage in conversations with a robot rather than with a human, as highlighted in the introduction, the same could be presumed for tablets.

Furthermore, our results suggest that children with ASD present less RRB's when interacting with a tablet than they do when interacting with a robot. Our third hypothesis has thus been confirmed. As previously mentioned, studies show a positive correlation between RRB's and anxiousness. Our results could thus mean that children with ASD tend to feel more anxious in the presence of a robot than they do in presence of a tablet. This could be due to familiarity: indeed, 10 out of 11 participants have access to a tablet at home, according to the demographic questionnaire completed by their parents. In fact, the mere-exposure effect suggests that people tend to prefer elements with which they are more familiar (Zajonc, 1968). One can thus wonder whether different results would have been found if the subjects had multiple sessions with the robot prior to the experiment. In this way, the children would have had similar levels of familiarity with both devices and the results would have been more conclusive.

Moreover, our analysis suggests that children with ASD are as likely to pay attention to a story and memorize its content if the latter is told by a robot, as they are if it is told by a tablet. Indeed, no statistically significant difference was found between the number of correct answers in each condition, thus disproving our initial hypothesis regarding cognitive performan-ces. These results support our behavioral data, and potentially indicate that tablets could be as efficient as robots in education-nal and therapeutic contexts. Robots have in fact been shown to improve the social skills of children with ASD (Cabibihan et al., 2013), and although further longitudinal studies are needed, our results may imply the same for tablets. Indeed, the teaching of social skills such as turn taking and imitation all require memory and attention which have not been found to vary when comparing robots and tablets. These results are interesting as tablets are more financially accessible than robots, and their use could easily be integrated in different types of therapy (Mechling, 2007).

As for the subjects' personal evaluation, our results show that the children preferred the story told by the tablet over the one told by the robot. Likewise, they would prefer to work with the tablet in the future; in fact, none of the participants chose the robot when answering this question. As explained previously, this could be due to their familiarity with the tablet. Our last two hypotheses regarding the children's evaluation of the interaction have thus been rejected. No statistically significant correla-tion was found between the participants' preferences and their age, nor with their SRS scores. This suggests that their age and social responsiveness skills had no impact on their preferred device.

In our study, the participants showed an equal willingness to communicate with the robot and with the tablet. Their cognitive performances were not particularly impacted by one device or the other. Although the subjects appeared to feel more at ease when interacting with the tablet, and expressed a general preference for the latter, these results are to be considered within the limitations of the study. Indeed, had the subjects also been familiarized with the robot prior to the experiment, the tendency to prefer the tablet might have been lower.

Furthermore, technical issues interfered with the experiment, such as the robot's face freezing or the tablet speaking in the wrong language for a few seconds. The children also appeared to be slightly distracted by the study conductors, especially during silent time. It is possible that their eye gazes would have been longer and/or more frequent if the experimenters had not been in the room. Additionally, the behavioral data could only be coded by a single team member, which means that it was not verified by the rest. This fact might also have slightly altered the results. Finally, as our sample size (N=11) was inferior to 15, we could only conduct non-parametric tests during the

statistical analysis (Field, 2009). In fact, two participants showed signs of distress and did not finish the study; one of them refused to participate altogether, whilst the other one interacted with the robot for a few seconds before running out of the room. They could therefore not be included in the results. We were also unable to study potential gender effects, as all participants who took part in the experiment were male.

Moreover, only two experimenters were present during the experiment itself. When conducting the study, the team quickly came to the conclusion that a group of three people would have been more adequate. In fact, the addition of another experimenter could have guaranteed a more precise documentation. In this scenario, a study conductor would have focused on technical matters, whilst the second one would have concentrated on the interaction with the children. Meanwhile, the third person would have documented the interactions in detail. In this way, confounding variables could have been noticed earlier.

It is also important to note that this study touches upon the attitudes of children with ASD towards robot and tablet storytellers, but not upon the efficiency of those devices in terms of behavioral therapy. Our findings concerning attention and memory could however be applied in educational contexts which require memorization and attention.

In conclusion, our initial results show no significant differences between the partici-pants' performances in each condition. In fact, the children generally felt more at ease with the tablet; most even stated that they preferred the latter over the robot. However, longitudinal studies with larger and more diverse samples would be necessary in order to draw any conclu-sions. These samples could include neuro-typical control groups and/or subjects with other neurodevelopmental disabilities. Considering the lower price and higher accessibility of tablets, it would in fact be very interesting to continue researching the efficiency of tablets in educational and therapeutic contexts. If results similar to ours were to be found in the future, Robot Assisted Therapy could start including the use of tablets and potentially become more accessible to the autistic population, both in professional institutions and at home. It would also be necessary to study the benefits and disadvantages of each device, in order to adapt their use to the children's individual needs and goals.

References

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.) https://doi.org/10.1176/appi .books.9 780890425596

Argyle, M. (1972). Non-verbal communication in human social interaction, *R. A. Hinde (Ed.) Non-verbal communication*. Oxford, England: Cambridge University Press.

Baird, G., Simonoff, E., Pickles, A., Chandler, S., Loucas, T., Meldrum, D. & Charman, T. (2006). Prevalence of disorders of the autism spectrum in a population cohort of children in South Thames: The Special Needs and Autism Project (SNAP). *The Lancet (British Editi-on)*, 368(9531), pp. 210 - 215. https://doiorg.proxy.bnl.lu/10.1016/S0140676(06)6 9041-7

Baribeau, D. A., Vigod, S., Pullenayegum, E., Kerns, C. M., Mirenda, P., Smith, I. M. & Szatmari, P. (2020). Repetitive behavior severity as an early indicator of risk for elevated anxiety symptoms in autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry,* 59(7), pp. 890 - 899. https://doiorg.proxy.bnl.lu/10.1016 /j.jaac.2019.08.478

Baron-Cohen, S., Scott, F. J., Allison, C., Williams, J., Bolton, P., Matthews, F. E., & Brayne, C. (2009). Prevalence of autism-spectrum conditions: UK school based population study. *The British journal of*

psychiatry, 194(6), pp. 500 - 509. doi: 10.1192/bjp.bp.108.059345

Bodfish, J. W., Symons, F. J., Parker, D. E., & Lewis, M. H. (2000). Varieties of repetitive behavior in autism: Compari-sons to mental retardation. *Journal of autism and developmental disorders*, 30(3), pp. 237 - 243. https://doi.org/10. 1023 /A:1005596502855

Cabibihan, J., Javed, H., Ang, M. & Aljunied, S. M. (2013). Why Robots? A Survey on the Roles and Benefits of Social Robots in the Therapy of Children with Au-tism. *International Journal of Social Robotics, 5*(4), pp. 593 - 618. https:// doiorg.proxy.bnl.lu/10.1007/s12369013-0202-2

Center for Disease Control and Prevention (2018). *Prevalence and Characteristics of Autism Spectrum Disorder Among Children Aged 8 Years. Autism and Developmental Disabilities Monitoring Network*, https://www.cdc.gov/mmwr/volumes/70/ss/pdfs/ss7011a1-H.pdf

Constantino, J. N., & Gruber, C. P. (2012). *Social Responsiveness Scale* Second Edition (SRS-2). Torrance, CA: Western Psychological Services

Costa, A., Charpiot, L., Lera, F., Ziafati, P., Nazarikhorram, A., Van der Torre, L., & Steffgen, G.. (2018). A Comparison between a Person and a Robot in the Attention, Imitation, and Repetitive and Stereotypical Behaviors of Children with Autism Spectrum Disorder. *Open Repository and Bibliography.*

Costa, T. Schweich, L. Charpiot, G. Steffgen et al. (2018). Attitudes of Children with Autism towards Robots: An Exploratory Study. Presented at *Interaction Design and Children (IDC-CRI2018) Workshop.* https://doi.org/10.1111/j.14 69 - 7610.19 96.tb01386.x

Dissanayake, C., & Crossley, S. A. (1996). Proximity and sociable behaviours in autism: Evidence for attachment. *Journal of child psychology and psy-chiatry*, 37(2), pp. 149 - 156.

Dunlap, G., Dyer, K., & Koegel, R. L. (1983). Autistic self-stimulation and intertrial interval duration. *American journal of mental deficiency,* 88(2), pp. 194 - 202.

Durand, V. M., & Carr, E. G. (1987). Social influences on “self-stimulatory” behavior: Analysis and treatment application. *Journal of Applied Behavior Analysis,* 20(2), pp. 119-132. https:// doi.org/10.1901/jaba.1987.20-119

Edwards. (2018, May). *Mix it up Monday: Consider the novelty effect*. https://edwardsvoice.wordpress.com/2018/05/07/mix-i t-up-monday-consider-the-novelty-effect/

Escobedo, L., Nguyen, D. H., Boyd, L., Hirano, S., Rangel, A., Garcia-Rosas, D., & Hayes, G. (2012, May). MOSOCO: a mobile assistive tool to support children with autism practicing social skills in real-life situations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 2589 - 2598

Fage, C., Consel, C. Y., Balland, E., Etchegoyhen, K., Amestoy, A., Bouvard, M., & Sauzéon, H. (2018). Tablet apps to support first school inclusion of children with autism spectrum disorders (ASD) in mainstream classrooms: A pilot study (2020). *Frontiers in Psychology*, 9. https://doi.org/10.3389/fpsyg.2018.02020

Feil-Seifer, D, & Matarić, Maja J. (2011). Socially Assistive Robotics. *IEEE Robotics & Automation Magazine*, 18(1), pp. 24 - 31. doi: 10.1109/MRA.2010.940150

Field, A. (2009) *Discovering Statistics Using SPSS*. 3rd Edition, Sage Publications Ltd., London.

Fombonne, Eric. (2009). Epidemiology of pervasive developmental disorders. Pe*diatric Research*, 65(6), pp. 591 - 598.

Gabriels, R. L., Cuccaro, M. L., Hill, D. E., Ivers, B. J., & Goldson, E. (2005). Repetitive behaviors in autism: Relationships with associated clinical features. *Research in developmental disabilities*, 26(2), pp. 169 - 181. https://doi.org/10.1016/j.ridd. 2004.05.003

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, 3(11), pp. 419 - 429. https://doi.org/10.1016/S1364-6613(99)01397-2.

Ho, S., Foulsham, T., & Kingstone, A. (2015). Speaking and listening with the eyes: gaze signaling during dyadic inter-actions. *PloS one*, 10(8), e0136905

Joyce, C., Honey, E., Leekam, S. R., Barrett, S. L., & Rodgers, J. (2017). Anxiety, intolerance of uncertainty and restricted and repetitive behaviour: Insights directly from young people with ASD. *Journal of Autism and Developmental Disorders*, 47(12), pp. 3789 - 3802.

Kose-Bagci H., Ferrari E., Dautenhahn K., Sverre D. S. & Nehaniv C. L. (2009). Effects of Embodiment and Gestures on Social Interaction in Drumming Games with a Humanoid Robot. *Advanced Robotics*, 23(14), pp. 1951 - 1996.

Lai, M.-C., Lombardo, M.V., & Baron-Cohen, S. (2014). Autism. *The Lancet (British Edition)*, 383(9920), pp. 896 - 910.

Lee, S., Odom, S. L., & Loftin, R. (2007). Social engagement with peers and stereotypic behavior of children with autism. *Journal of Positive Behavior Interven-tions*, 9(2), pp. 67 -79. https://doi.org/ 10.1177/10983007070090020401

Lewis, M. H., & Bodfish, J. W. (1998). Repetitive behavior disorders in autism. *Mental retardation and developmental disabi-lities research reviews*, 4(2), pp. 80 - 89. https://doi.org/10.1002/(SICI)1098-2779(1998)4:2<80::AID-MRDD4>3.0.CO;2 - 0

Loftin, R. L., Odom, S. L., & Lantz, J. F. (2008). Social interaction and repetitive motor behaviors. *Journal of Autism and Developmental Disorders,* 38(6), pp. 1124 - 1135. https:// doi.org/10.1007/ s10803-007-0499-5

Lord, C., Elsabbagh, M., Baird, G., & Veenstra Vanderweele, J. (2018). Autism spectrum disorder. *The Lancet (British Edition)*,392(10146), pp. 508 - 520. https://doi-org.proxy.bnl.lu/10.1016/S0140-6736(18)31129-2

Louie, Korneder, Abbas, Pawluk. (2020). A study on an applied behaviour analysis-based robot-mediated listening comprehension intervention for ASD 2020. *Paladyn, Journal of Behavioural Robotics*, vol. 12, no. 1, pp. 31 - 46. https:// doi.org/10.1515/pjbr-2021-0005

Lux AI. (2019, May). *Why do children with autism learn better from robots* https://luxai.com/blog/why-children-wit h-autism-learn-better-from-robots/

Mechling, L. C. (2007). Assistive technology as a self-management tool for prompting students with intellectual disabilities to initiate and complete daily tasks: a literature review. *Education and Trai-ning in Developmental Disabiliti-es,*42(3), pp. 252 - 269. http://www.jst or.org/stable/23879621

Melo, F. S., Sardinha, A., Belo, D., Couto, M., Farias, A., Ventura, R. (2019). Project INSIDE: Towards autonomous semi-unstructured human–robot social interaction in autism therapy. *Artificial Intelligence in Medicine*, 96, pp. 198 - 216. https://doi-org.proxy.bnl.lu/10.10 16/j.artmed.2018.12.003

Michaelis, M. Gombolay, M., De Graaf, T. (2021). Novelty Effects in *Human-Robot Interaction.* Frontiersin.

O'Haire, M.E. (2012). Animal-Assisted Intervention for Autism Spectrum Disorder: A Systematic Literature Review. *Journal of Autism and Developmental Disorders*, 43(7), pp. 1606 - 1622. doi: 10.1007/s10803-012-1707-5

Pino, O., Palestra, G., Trevino, R. et al. (2020). The Humanoid Robot NAO as Trainer in a Memory Program for Elderly People with Mild Cognitive Impairment. *Int J of Soc Robotics* 12, pp. 21 - 33. https://doi-org.proxy.bnl.lu/10. 1007/s12369-019-00533-y

Pop, C. A., Simut, R. E., Pintea, S., Saldien, J., Rusu, A. S., Vanderfaeillie, J. & Vanderborght, B. (2013). Social robots vs. computer display: Does the way social stories are delivered make a difference for their effectiveness on ASD children?. *Journal of Educational Computing Research*, 49(3), pp. 381 – 401 http://dx.doi.org/10.2190/EC.49.3.f

Robins, B., Dautenhahn, K., & Dubowski, J. (2006). Does appearance matter in the interaction of children with autism with a humanoid robot?. *Interaction studies*, 7(3), pp. 479 - 512. https://doi.org/10.1075/is.7.3.16rob

Saad, J. Broekens, M. Neerincx, K.V. Hindriks. (2020). Enthusiastic Robots Make Better Contact. Presented at *EEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*

Scassellati, B, & Admoni, H. & Matarić, M. (2012). Robots for Use in Autism Research. *Annual review of biomedical engineering*. 14. pp. 275 - 294. 10.1146/annurev-bioeng-071811-1500 36.

Srinivasan, S.M., Eigsti, I.-M., Neelly, L., & Bhat, A.N. (2016). The effects of embodied rhythm and robotic interventions on the spontaneous and responsive social attention patterns of children with autism spectrum disorder (ASD): A pilot randomized controlled trial. *Research in Autism Spectrum Disorders*, 27, pp. 54 - 72. https://doi-org.proxy.bnl.lu/10. 1016/j.rasd.2016.01.004

Szymona, B., Maciejewski, M., Karpiński, R., Jonak, K. Radzikowska-Büchner, E., Niderla, K. & Prokopiak, A. (2021). Robot-Assisted Autism Therapy (RA AT). Criteria and Types of Experiments Using Anthropomorphic and Zoomor-phic Robots. Review of the Re-search. in *Sensors* (Basel, Switzer-land), 21(11), p. 3720. https://doi.org/ 10.3390/s21113720

Thill, S., Pop, C. A., Belpaeme, T., Ziemke, T., & Vanderborght, B. (2012). Robot-assisted therapy for autism spectrum disorders with (partially) autonomous control: Challenges and outlook. *Paladyn*, 3(4), pp. 209 - 217. https://doi-org.proxy.bnl.lu/10.2478/s13 230-013-0107-7

Turner, M. (1999). Repetitive behaviour in autism: A review of psychological research. *The Journal of Child Psychology and Psychiatry and Allied Disciplines,* 40(6), pp. 839 - 849. https:// doi-org.proxy.bnl.lu/10.2478/s1323001 3-0107-7

Vanderborght , B. (2012). Robot assisted therapy for autism spectrum disorders with (partially) autonomous control: Challenges and outlook. *Paladyn*, 3(4), pp. 209 - 217.

Waterhouse, L. (2013). *Rethinking Autism : Variation and Complexity* (1st ed.). London ; Waltham, MA: Academic Press.

The Impact of Hemispheric Laterality on Interoceptive Processing

Meggie Barnabo, Mareike Boos, Jessica Goergen, Franziska Leufgen, Emily Schramm, Joy Steinmetzer
Supervision: Sam Bernard (M.Sc., Doctoral Researcher)

This study aimed at gaining a deeper insight on Hemispheric Laterality and its impact on interoception. We investigated a potential link between being left- or right-handed and the perception of one's own body. A Heartbeat Counting Task was administered, in which participants reported the number of their perceived heartbeats in comparison to those measured with an ECG recording. Afterwards, a Time Estimation Task, serving as the control condition, was administered. Given the results of previous research, we expected to find a significant difference in Interoceptive Accuracy when controlling for Hemispheric Laterality, which could however not be found in our sample of $N = 42$ participants. Nevertheless, we found a difference between female left- and right-handers at trend level, showing that female left-handers achieve highest Interoceptive Accuracy scores. In addition, investigating the gender difference within right-handed participants only, another result at trend-level was found. Moreover, a significant difference between time conditions was revealed, indicating that Interoceptive Accuracy scores were highest performing the task during the later afternoon.

Introduction

Interoception is what we can describe as the conscious and unconscious perception of our own body, specifically of the body's physical condition and our emotional state. It has to be differentiated from exteroception, describing the perception of one's environment. During the last years, the importance of Interoception in psychological research has grown, as it holds a "fundamental role [...] in human con- sciousness" and has been found to be able to influence human behavior via motivational processes (Craig, 2003). Implications of a dysfunction in Interoceptive processing leading to mental health issues such as eating disorders or mood and anxiety disorders have been made, which make this topic clinically relevant. (Khalsa et al., 2018). Recent psychophysiological findings have revealed that the In- teroceptive cortex, represented in the right an- terior insula, receives sensory input from the entire body and not just from the viscera as assumed previously. It has therefore been described as the "long-missing afferent complement of the efferent autonomic nervous system (ANS)" (Craig, 2002). As Interoceptive processes interact with cognition and emotion, there are interindividual differences between humans concerning the ability to perceive these inner bodily feelings as well as the characteristics of the Interoceptive feelings themselves. It can be differentiated between Interoceptive Accuracy, which describes the objective Interoceptive ability of- ten indicated by behavioral performances, and Interoceptive Sensibility, the subjective beliefs in one's Interoceptive ability (Garfinkel et al., 2015).

In order to quantify the construct of Interoceptive Accuracy, Schandry et al. have established a paradigm frequently used in experimental research: In this task, test subjects have to count their own heartbeat in different time intervals without external help, such as feeling their own pulse. Their subjective reports will then be compared to the actual number of heartbeats in the respective time intervals, resulting in an objective measure of their interoceptive ability. It is important to stress that subjects classified as good perceivers of their own heartbeat showed higher levels of state anxiety and higher levels of emotional lability as a personality trait in the original study (Schandry, 1981). However, these findings could not be

supported in a replication of Schandry's study by Montgomery and Jones, resulting in the assumption that factors like the respiration rate as a result of general autono- mic arousal rather than state or trait anxiety – making obesity another indicator of Interocep- tive Accuracy - could possibly be influencing interoceptive ability (Montgomery & Jones, 1984). However, it should not be argued that emotional experience ought to be considered when administering Schandry's heartbeat counting task, as emotion is still coupled to in- teroceptive processing. As a measure of In- teroceptive Sensibility, the subject's con- fidence ratings in their own answer after each interval compared to the actual Interoceptive Accuracy can be integrated in the Schandry task. Previous findings suggest that there is almost no correspondence between Interocep- tive Sensibility and Interoceptive Accuracy (Garfinkel et al., 2015). Hence, regarding the relationship of Interoceptive Accuracy and Sensibility as well as the key factors influen- cing Interoceptive ability, the aim of this study is to investigate this topic further.

Our research question is based on various findings with reference to the influence of Hemispheric Laterality on interoception. Studies making use of the heartbeat detection paradigm as an operationalization of Interoceptive Accuracy have shown that right hemisphere preference was significantly related to the performance in the heartbeat detection task (Montgomery & Jones, 1984). These findings may be explained by a possible relation of in- teroceptive ability to cortical events in the right hemisphere (Katkin et al., 1991). In order to differentiate between right and left hemi- spheric preference, conjugate level eye mo- vements were evaluated. This paradigm how- ever has been quite controversial, as the direction of eye movement can be influenced by other factors and distractors (Hantas et al., 1984).

In our study, we therefore want to review and possibly verify the findings on Interoceptive Accuracy being influenced by right hemi- spheric preference, the latter being operationalized by differentiating between left- handedness and right-handedness instead of conjugate level eye movement measurement. We will administer Schandry's heartbeat de- tection task and compare Interoceptive Ac- curacy to Interoceptive Sensibility, measured by the participant's self-reported confidence in their answer, and further investigate the relati- onship between interoceptive ability and state anxiety, as previous findings on these aspects have resulted in opposing implications. Furthermore, a possible gender difference will be investigated, as Montgomery and Jones have suggested that men could show better performance in a heartbeat detection task due to a stronger lateralization.

Hence, we postulate the following hypotheses:

1. There is a significant difference in Interocepti- ve Accuracy when controlling for Hemispheric Laterality (Katkin et al., 1991)
2. A) There is a significant difference between left-handed and right-handed people concern- ing their score in the heartbeat discrimination task which can be considered a measure for Interoceptive Accuracy (Montgomery & Jones, 1984)
   B) Left-handed people will score higher in In- teroceptive Accuracy than right-handed peop- le (Hantas et al., 1984)
3. There is no significant difference in performing the Time Estimation Task between left-handed and right-handed people.

Methodology

Materials/Measures

To begin with, we used a proprietary self- report questionnaire to assess basic demo- graphic, medical and physiological markers, including age, gender, history of past or present illness, medication, sleep quality and quantity, as well as height and weight (used to evaluate the BMI).

In order to confirm the participant's reported handedness, the 10-item Edinburgh handed- ness Inventory (Oldfield, 1971) was used. A variety of different "everyday tasks'' such as "with which hand do you normally use the broom when swiping the floor?'' was given and the participant then had

to mark out which hand is more commonly used to realize the task. In the end, all ticks per hand were added and the sum which was greater, indicated the participants handedness.
In addition, we used the State-Trait-Anxiety Inventory (Spielberger, Gorsuch, Lushene,Vagg, & Jacobs, 1983) which has forty items in total. The inventory consists of the STAI-T Questionnaire, which participants had to fill out before completing both the Heartbeat Counting and Time Estimation Task. Partici- pants could choose which item corresponded to their actual well-being by using a 4-point Li- kert scale ranging from 1 (almost never) to 4 (very often). Items like “I am satisfied”, “I am happy” and “I feel like, I want to cry” are examples of statements stated in the Inventory.
At the end of the experiment, consequently after completing the task, participants then had to fill out the second part of the inventory, the STAI-S Questionnaire. Here, typical statements are “I am nervous”, “I am content” and “I am calm”.

For the measurement of the participant’s actual heart rate, the Polar Watch RS800CX was used, to which participant’s heartbeats were transferred wirelessly from two ECG electrodes on their chest. Using the software PolarTrainer5 allowed us to extract the recorded number of heartbeats during each interval and to calculate an Interoceptive Accuracy score.

Participants

Altogether, *N* = 42 participants took part in the study. The average age was *M* = 25.93 (*SD* = 10.24). Participants were aged between 18 and 55 years. Concerning the gender distribu- tion, 47.6 % of the participants were male and 52.4 % were female. In total, 50 % were left- handed (*N*=21) whereas 50 % (*N*=21) indica- ted to be right-handed.
While 28.4 % indicated that they work, 71.4 % reported they do not. Finally, 69 % of the participants indicated to be students.

Study procedure

We divided our sample into two groups, according to their handedness. During the experiment, all participants did both the heartbeat counting task (experimental condition) first as well as the Time Estimation Task (control condition) after in order to avoid priming effects. Each task contained a practice trial as well as six trials of 25, 35, 45, 55, 65 and 75 seconds. The order was assigned randomly to each participant. Participants were either invited to a dedicated experimental room on the University of Luxembourg’s Campus for the duration of the study (*N* = 16), however, due to practical reasons such as a limited testing period of four weeks, most participants were tested at their homes (*N* = 26). Firstly, they received an information sheet with general information on the study and about the nature of the task to follow, without being told explicitly the object that is in focus of the estimation or counting task.

Afterwards, they were asked to fill out all questionnaires mentioned above, except for the STAI-S, which they were asked to answer after the actual experiment. In this part of the study, their reported hemispheric preference was confirmed using the Edinburgh Scale.

Participants were then sat in front of a laptop, the experimenter retreating to the background in order for the participants to not feel observed. Firstly, participants were told to relax during a five-minute resting period, so that their usual resting heart rate could be reached and stabilized. They were then introduced to the Schandry Heartbeat Counting Task, in which they had to perceive and count their own heartbeat in the six given intervals. After each trial, they were told to rate their con- fidence in their own answer on a scale of zero to eight. The Heartbeat Counting Task was then immediately followed by the Time Estima- tion Task, in which participants were asked to indicate how many seconds had passed in the given intervals, and again report their con- fidence in their answer on the same scale.
After the experimental session of about forty minutes in total was over, participants were

reimbursed for their time by being given an incentive of 10€ gift voucher and students of the University of Luxembourg could choose to be accredited participation hours for their respective courses additionally.

Analysis

Our analysis was conducted using the statistical software IBM SPSS 27. Due to corrupted ECG recordings, two participants of the original sample of $N = 44$ participants had to be excluded.

The State-Trait-Anxiety Inventory Questionnaires STAI-T and STAI-S were evaluated based on raw values, each serving as a new state- and trait anxiety variable. Two new variables were computed, one being the In- teroceptive Accuracy score and the other being the Time Estimation Accuracy score, u- sing a commonly used formula: 1- 1/n ∑ |recorded heartbeats-reported heart- beats|/recorded heartbeats, *n* being the num- ber of trials, in our case $n = 6$ (Ainley et al., 2020). As the instructions were not correctly understood, leading to a deviation of more than two Standard Deviations ($SD = .24$) from the mean ($M = .64$), we had to exclude two more participants for the Heartbeat Counting Task, which resulted in new total of $N = 37$ participants for the Heartbeat Counting Task. The same issue occurred in the Time Estima- tion Task ($M = .81$, $SD = .12$), leading to a total of $N = 41$ for the analyses of the Time Estima- tion condition. In order to report the Interocep- tive Sensitivity for each of the two tasks, we computed two more variables out of the mean values taken from the Likert scales.

Before conducting any analyses, the assumption of normality was investigated with a Kolmogorov-Smirnov Test for Normality as well as a Shapiro-Wilk Test for Normality. How- ever, as the sample size was $N > 20$ ($N = 42$) and therefore part of the “central limit theorem” (Van den Berg, R. G., 2020) it is to conclude that the sample is normally distributed.
First, a frequency analysis of the sample was conducted, and in order to test our hypotheses, we conducted several correlative analyses as well as t-Tests and Analyses of Variance (ANOVA).

Results

The results of the correlation analysis show that Interoceptive Accuracy and Interoceptive Sensibility relations are not significantly associated, $r = .09$, $p = .59$. These results align with findings of Garfinkel et al, 2015. As expected, Time Estimation Accuracy and Time Estimation Sensibility are also not significantly correlated, $r = -.03$, $p = .87$.

To compare the means of right-handed ($M = .65$, $SD = .23$) and left-handed participants ($M = .70$, $SD = .18$) with regards to their Interoceptive Accuracy score, a t-Test for independent samples was conducted. As bar chart 1 shows, it revealed that there is no significant difference within handedness $t(35) = -.69$, $p = .50$, $d = .20$ which contradicts our main hypothesis of left-handed people performing better at the Heartbeat Counting Task.

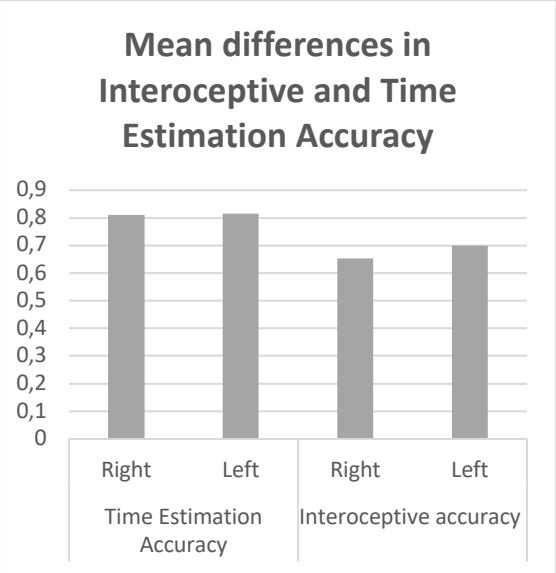
However, breaking down the heterogeneity of the sample and investigating the influence of handedness on Interoceptive Accuracy only in women ($N = 19$), we found a difference between female left- and right-handers at trend level, $t(17) = -2.05$, $p = .056$, $d = .24$. The outcomes are underlined in bar chart 2, which shows that within the highest scoring group in Interoceptive Accuracy, only left-handed females were represented. Considering the difference that was found in males ($N = 18$) which was not significant at all, $t(16) = .35$, $p = .73$, $d = .23$, these findings could indicate that in a larger sample, a possible gender difference exists, hence the effect of laterality on interoception is overpowered by the heterogeneity of the complete sample and therefore only appears within females.

To compare the means of right-handed ($M = .65$, $SD = .23$) and left-handed participants ($M = .70$, $SD = .18$) with regards to their Time Estimation Accuracy score, another t-Test for independent samples was conducted. It revealed that there is no significant difference within

handedness, $t(39) = -.15$, $p = .88$, $d = .12$, which confirms the third hypothesis stating that right- and left-handed people do not differ in regards to their time estimation ability. This was to be expected from the task serving as the control condition.

Further, we investigated a possible gender difference on Interoceptive Accuracy using a t-Test for independent samples. Again, no significant difference between female and male participants was found, $t(35) = -1.6$, $p = .12$, $d = .20$. This contradicts the proposed hypothesis by Montgomery and Jones, 1984 stating that men could outscore women in the Heartbeat Counting Task due to a stronger lateralization.

However, investigating the gender difference within right-handed participants only, breaking down the heterogeneity of the sample once again, a difference at trend level between female and male participants was found, $t(17) = -2.10$, $p = .053$, $d = .21$ considering that the Levene's Test for Equality of Variances was significant and equal variances could not be assumed ($p = .01$). These findings are again of special interest given that there was no significant difference between gender groups in left-handed people $t(16) = -.19$, $p = .85$, $d = .18$. Once more, these results stipulate that a possible gender difference of a larger sample is overpowered by the heterogeneity of the total sample.

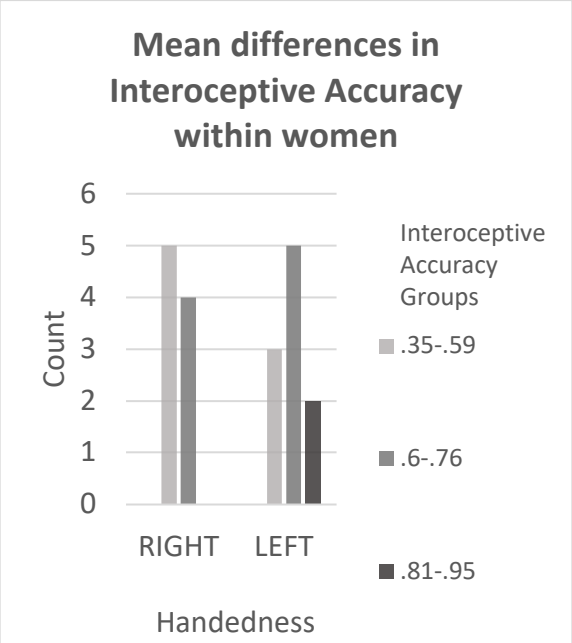


Bar chart 1: Mean differences in interoceptive and Time Estimation Accuracy

No significant difference between female and male participants was found regarding the Time Estimation Accuracy score $t(39) = -.64$, $p = .524$, $d = .12$, showing that men and women do not differ significantly in their time estimation ability.

A 2 x 2 Univariate ANOVA was conducted in order to confirm the t-Test results and furthermore probe a possible interaction effect on handedness and gender. Neither a significant main effect of handedness was found, $F(1, 33) = .65$, $p = .43$, $\eta^2 = .25$ nor a significant interaction effect of the two factors was found, $F(1, 33) = 1.95$, $p = .17$, $\eta^2 = .06$. Although there was no significant main effect of gender on Interoceptive Accuracy to be found, $F(1, 33) = 2.7$, $p = .11$, the effect size of this factor on the dependent variable was of medium size, $\eta^2 = .58$. This supports our findings on a possible gender difference being overpowered by the heterogeneity of the total sample within right-handed people.

Additionally, based on propositions in literature (mentioned in the introductory section), we conducted various correlative analyses to investigate the influence of different controlled extraneous variables on Interoceptive Accuracy, which all showed



Bar chart 2: Mean differences in Interoceptive Accuracy within women

non-significant outcomes: age (r = .15, p = .37), BMI (r = .03, p = .87), quality of sleep (r =.15, p = .37), duration of sleep (r = -.06, p = .71), medication (r = -.18, p = .30) trait-anxiety (r = -.02, p = .12) and state-anxiety (r = -.18, p = .08). The results regarding state anxiety align with findings by Montgomery and Jones, 1984, who could also not support Schandry's hypothesis that good heartbeat perceivers showed higher state anxiety in the first place. In addition to the findings on state-anxiety, we could now show that anxiety as a personality trait does not seem to have an influence on Interoceptive Accuracy either.

To investigate the influence of testing time on Interoceptive Accuracy, we differentiated between three time groups, ranging from 9:00 – 12:00 o'clock, 12:00 – 16:00 o'clock and from 16:00 o'clock onwards. Results of a correlation analysis revealed a significant linear relationship between time groups and Interoceptive Accuracy, r = .55, p = .000.

A One-Way ANOVA revealed that there are significant differences between the time groups with regards to the Interoceptive Accuracy score, F (2, 34) = 8.24, p = .001, η^2 = .33. A post-hoc test (Scheffe) revealed significant differences between time conditions, specifically between groups 3 and 1 as well as between groups 3 and 2, as the mean score of people who completed the task after 16:00 o'clock (M = .80, SD = .15) was significantly higher than the mean score of people who completed the task between 9:00 and 12:00 o'clock (M = .53, SD = .18, p = .008) as well as the mean score of people who completed the task between 12:00 and 16:00 o'clock (M = .60, SD = .19, p = .008). Table 1 shows the impact of these results: If Interoceptive Accuracy scores are divided into three different groups (.35 - .59; .60 - .76; .81 - .95), ten out of twelve people in the best-performing group (a score of .81 onwards) completed the task later than 16:00 o'clock.

			Time Groups			
			9:00-12:00	12:00-16:00	From 16:00	Total
Interoceptive Accuracy groups	.35 - .59	Count	4	8	1	13
		% within time groups	66,7%	53,3%	6,3%	35,1%
	.6 - .76	Count	2	5	5	12
		% within time groups	33,3%	33,3%	31,3%	32,4%
	.81 - .95	Count	0	2	10	12
		% within time groups	0,0%	13,3%	62,5%	32,4%
Total		Count	6	15	16	37
		% within time groups	100,00%	100,00%	100,00%	100,00%

Table 1: Interoceptive Accuracy groups * time groups Crosstabulation

Conclusion and Discussion

Summary

Based on the conducted analyses and in the present sample, we could not support the hypothesis of an impact of one specific hemispheric dominance on Interoceptive Accuracy per se. Nevertheless, interesting effects of gender at trend level were found, proposing that an effect of laterality could only be present within women, so that left-handed women reach higher Interoceptive Accuracy scores on the Schandry Heartbeat Counting Task than right-handed women, which can reversely be seen as an indicator of our second hypothesis of left-handedness favoring Interoceptive Accuracy. However, this remains to be investigated further. The same applies to the second trend-level effect that could be found regarding a gender effect within right-handed people.
The main aim of our study was to further investigate on findings provided in previous research and contribute to a clarification of contradictory results of the studies on the relationship between Hemispheric Laterality and interoceptive processing that had been done already. Furthermore, the results implicated interesting aspects to consider in future research on this topic, given the example of a significant effect of testing time on Interoceptive Accuracy.

Limitations and Outlook

Given that we were unable to find statistical support for our main hypothesis in this sample, the question arises to what extent this study is of scientific relevance. Replications of Schandry's original study have shown contradictory results, for instance the influence of state-anxiety on interoceptive ability (Montgomery & Jones, 1984). Therefore, every attempt to provide further research in the field of hemispheric dominance and interoceptive ability is important in order to gain a deeper insight on the topic.
This study specifically tried to control for many extraneous variables such as Body Mass Index or testing time on the relationship between handedness and Interoceptive Accuracy, which have also been neglected before.
Moreover, the sample consisted of both female and male participants, unlike many studies before, in which only men were recruited.
We made use of the Schandry Heartbeat Counting Task, which is an established and economical method to measure interoceptive ability. Although there has been an accumulation of criticisms on the task recently,

heretofore there is no alternative method available which delivers better results. In Schandry's original study, analyses were based on an allocation of all participants to two groups of "good perceivers" and "poor perceivers" (Schandry, 1981). However, this turned out to be problematic regarding the comparison of the task and its results to other Heartbeat Detection Tasks (Knoll & Hodapp, 1992). For this reason and due to our rather small sample size, the analyses in this study were based on the Interoceptive Accuracy score scale as a whole, leading to an avoidance of focusing on extreme groups.

A post-hoc G*Power analysis revealed that in order to be able to detect a significant effect (d = .50, α = .05) with a statistical power of .80, a total sample size of N = 128 is needed.

In general, our study took place in different locations and at different times of the day. While the significant effect of task completion time revealed interesting new implications and remains to be further investigated, the lack of standardization of testing in one specific location could have been a possible reason for the lack of significant results, as it has been shown before that interoceptive tasks are very sensitive to the testing context (Ainley et al., 2020). In order to achieve a more reliable and stable Interoceptive Accuracy score, future study designs could include various testing times for each participant.

It must also be taken into account that this study took place during the global COVID-19 pandemic, which could have influenced the participant's state-anxiety scores, that showed a rather high correlation to Interoceptive Accuracy scores compared to the relationship between Interoceptive Accuracy and other extraneous variables.

In addition, it should also be considered that interoceptive processing is only partly and to varying degrees conscious, the Heartbeat Counting Task is therefore only able to capture the self-reported conscious component of interoceptive perception. To perceive all processes, an EEG-recording providing heartbeat-evoked potentials would be preferable. While interoceptive ability is therefore not measured as a whole with the Schandry task, the advantage of the method being non-invasive outweighs this factor to a certain degree (Khalsa et al., 2018).

The use of the Polar Watch RS800CX represents a further methodological challenge, as it is not possible to check whether participant's heartbeats are correctly reported over the entire recording span or whether artefacts occur (Ainley et al., 2020). For each participant, recordings had to be checked thoroughly after the experiment. This led to the exclusion of three participants for the Heartbeat Counting Task analyses in this study. Given the context of this study being a training in practical research, the Polar Watch is the most economical method to save time and resources, however, we would suggest using a proper ECG recording in future studies.

As indicated above, more criticism on Schandry's Heartbeat Counting Task has arisen frequently. In general, non-interoceptive processes presumably influence interoceptive processing (Desmedt et al., 2018). It could for instance be possible that people use their subjective sense of time as an orientation when being asked to report their counted heartbeats, resulting in the Interoceptive Accuracy score not depending on perceived heartbeats but estimated heartbeats (Knoll & Hodapp, 1992). This assumption is supported by findings suggesting that higher intelligence, and therefore a better knowledge on a normal resting heartbeat, favors better heartbeat estimation ability (Murphy et al., 2018). Additionally, it has been shown that people tend to underestimate their own heartbeat (Ainley et al., 2020).

A last point worth mentioning is that Schandry's Heartbeat Counting Task is probably not able to measure interoceptive ability as an overarching skill, especially regarding the measurement of Interoceptive Sensitivity, as the heartbeat counting and self-report of confidence in one's answer after each trial is not a sufficient measure of all existing sub-facets of Interoceptive Sensitivity (Ring & Brener, 2018).

Nevertheless, if these limitations are considered, Schandry's task remains one of the most reliable measure for Interoceptive Accuracy in experimental studies that exists at the moment.

For future research, we also suggest to include other interoceptive systems into the measurement of interoceptive processing other than the cardiovascular system, such as the gastrointestinal, thermoregulatory or nociceptive system among others (Khalsa et al., 2018), as this would ensure a more complex measurement of one's Interoceptive Ability.
Concerning control variables, as the interaction of interoception with cognition and emotion has been proven (Garfinkel et al., 2015), it could be possible that other emotional proceses apart from anxiety and stress have an impact on interoceptive processing. Therefore, it may be reasonable to control for more emotional states, such as joyful excitement, pleasure or other emotions that could also explain a state of heightened physical vigilance.

Regarding the results of this study, we would also suggest to investigate the influence of testing time on Interoceptive Accuracy further. Moreover, more studies on a possible gender difference regarding Interoceptive Accuracy and Hemispheric Laterality are needed.

References

Ainley, V., Tsakiris, M., Pollatos, O., Schulz, A., & Herbert, B. M. (2020). Comment on "Zamariola et al. (2018), Interoceptive Accuracy Scores are Problematic: Evidence from Simple Bivariate Correlations"—The empirical data base, the conceptual reasoning and the analysis behind this statement are misconceived and do not support the authors' conclusions. *Biological Psychology*, *152*, 107870. https://doi.org/10.1016/j.biopsycho.2020.107870

Craig, A. (2003). Interoception: The sense of the physiological condition of the body. *Current Opinion in Neurobiology*, *13*(4), 500–505. https://doi.org/10.1016/S0959-4388(03)00090-4

Craig, A. D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews Neuroscience*, *3*(8), 655–666. https://doi.org/10.1038/nrn894

Desmedt, O., Luminet, O., & Corneille, O. (2018). The heartbeat counting task largely involves non-interoceptive processes: Evidence from both the original and an adapted counting task. *Biological Psychology*, *138*, 185–188. https://doi.org/10.1016/j.biopsycho.2018.09.004

Garfinkel, S. N., Seth, A. K., Barrett, A. B., Suzuki, K., & Critchley, H. D. (2015). Knowing your own heart: Distinguishing interoceptive accuracy from interoceptive awareness. *Biological Psychology*, *104*, 65–74. https://doi.org/10.1016/j.biopsycho.2014.11.004

Hantas, M. N., Katkin, E. S., & Reed, S. D. (1984). Cerebral Lateralization and Heartbeat Discrimination. *Psychophysiology*, *21*(3), 274–278. https://doi.org/10.1111/j.1469-8986.1984.tb02934.x

Katkin, E. S., Cestaro, V. L., & Weitkunat, R. (1991). Individual Differences in Cortical Evoked Potentials as a Function of Heartbeat Detection Ability. *International Journal of Neuroscience*, *61*(3–4), 269–276. https://doi.org/10.3109/00207459108990745

Khalsa, S. S., Adolphs, R., Cameron, O. G., Critchley, H. D., Davenport, P. W., Feinstein, J. S., Feusner, J. D., Garfinkel, S. N., Lane, R. D., Mehling, W. E., Meuret, A. E., Nemeroff, C. B., Oppenheimer, S., Petzschner, F. H., Pollatos, O., Rhudy, J. L., Schramm, L. P., Simmons, W. K., Stein, M. B., … Interoception Summit 2016 Participants. (2018). *Interoception and Mental Health: A Roadmap* [Application/pdf]. 13 p. https://doi.org/10.3929/ETHZ-B-000282635

Knoll, J. F., & Hodapp, V. (1992). A Comparison between Two Methods for Assessing

Heartbeat Perception. *Psychophysiology*, *29*(2), 218–222. https://doi.org/10.1111/j.1469-8986.1992.tb01689.x

Montgomery, W. A., & Jones, G. E. (1984). Laterality, Emotionality, and Heartbeat Perception. *Psychophysiology*, *21*(4), 459–465. https://doi.org/10.1111/j.1469-8986.1984.tb00227.x

Murphy, J., Millgate, E., Geary, H., Ichijo, E., Coll, M.-P., Brewer, R., Catmur, C., & Bird, G. (2018). Knowledge of resting heart rate mediates the relationship between intelligence and the heartbeat counting task. *Biological Psychology*, *133*, 1–3. https://doi.org/10.1016/j.biopsycho.2018.01.012

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Ring, C., & Brener, J. (2018). Heartbeat counting is unrelated to heartbeat detection: A comparison of methods to quantify interoception. *Psychophysiology*, *55*(9), e13084. https://doi.org/10.1111/psyp.13084

Schandry, R. (1981). Heart Beat Perception and Emotional Experience. *Psychophysiology*, *18*(4), 483–488. https://doi.org/10.1111/j.1469-8986.1981.tb02486.x

Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Van den Berg, R. G. (2020). *SPSS Kolmogorov-Smirnov Test for Normality*. SPSS Tutorials. https://www.spss-tutorials.com/spss-kolmogorov-smirnov-test-for-normality/

Do hormones matter? The influence of menstrual cycle phase and hormonal contraception on body image distortion and body (dis)satisfaction in adult women.

Danaé D. Lamy-Au-Rousseau, Isabella De Sousa Pereira, Laurie Henkes, Nicola Theis, Renata Esayan
Supervisor: M.Sc. Lynn Erpelding

Body image refers to the perception of and thoughts and feelings about the own body. Body image distortion relates to the misestimation of the own body size (perceptive body image component), while body dissatisfaction implies negative thoughts and feelings to- wards the body (cognitive-affective component). Although a key symptom of eating dis- orders, body image disturbance can also be observed in almost all healthy women. Fluc- tuations of sex hormones across the menstrual cycle change women's body perception, as well as thoughts and feelings towards their bodies. This study seeks to compare nat- urally cycling women in the perimenstrual phase, naturally cycling women in the intermenstrual phase, and women on hormonal birth control regarding their body size esti- mation and body dissatisfaction. Twenty-two healthy women performed both a metric body size estimation (BSE) task, as well as a depictive BSE task in VR to estimate their body size. Body dissatisfaction (BID) was measured with different self-report measures. Regarding the BSE-VR task, naturally cycling in the intermenstrual phase differed signif- icantly from both naturally cycling in the perimenstrual phase ($p = .001$, $\eta 2 = .566$) and women on hormonal birth control ($p = .008$, $\eta 2 = .566$), while no significant difference between naturally cycling in perimenstrual phase and women on hormonal birth control could be found. No significant group differences were found in the metric BSE task and body dissatisfaction scores. Hormone levels, thus, seem to have an influence on BSE, but only on estimation in a task from an allocentric reference frame (VR-BSE), while no group differences have been found for the BSE estimation task from an egocentric ref- erence frame (metric BSE task), which has already been shown in previous studies. Contrary to previous findings, no group differences in body dissatisfaction could be found. Strengths and limitations of the present study are discussed with a focus on future re- search

Introduction

“Body image describes the relation be- tween a human and his thoughts and feelings about his own body” (APA Dic- tionary of Psychology, n.d.). Body image has three different components. First the perceptive component which is the de- tection, estimation and identification of one’s own body size. It is defined as the accuracy of the individuals’ judgement of their size, shape and weight regarding their actual proportions. Second is the affective component, this mostly involves the feelings that individuals develop towards their bodies’ appearance and the (dis)satisfaction of one’s body. Lastly, the cognitive component mostly relies on the beliefs regarding body shape and appearance and the mental representation of one’s own body (Gaudio & Quattrocchi, 2012).

Body image distortion and body dissat- isfaction (BID) can lead to mental disor- ders, such as various eating disorders. The most affected eating disorders are anorexia nervosa (AN) and bulimia ner- vosa (BN). The key symptoms of AN are

nourishment restriction and being underweight, which causes the body weight to be lower than the normal BMI limit of greater than or equal to 18.5 kg/m^2 (Ad- ministration Substance Abuse and Men- tal Health Services, 2016; American Psychiatric Association, 2015; Kring et al., 2019). Most people with this

disorder show a severe fear of weight gain, and they seek to lower their body weight, which leads to further distortion of the perception of their own body (body im- age disorder). The main symptoms of BN are eating attacks with subsequent vomiting, which is also a symptom of the binge-purging subtype of AN. Affected people have repeated, frequent eating episodes in which they consume several thousand calories in a short period of time. However, they use counteracting or compensating measures, such as self-induced vomiting, to prevent weight gain. Often, their figure and weight greatly influence their self-worth and self-esteem (American Psychiatric As- sociation, 2015).

Surprisingly, body image distortion can be found in almost all healthy women (Fuentes et al., 2013). Previous studies have shown that girls and female ado- lescents and women of all ages outlined BID. However, other studies also indi- cate that BID differs in different age groups. (Frederick et al., 2006) suppose that 20% to 40% of women are dissatis- fied with their bodies and 10% to 30% of the men. Further studies investigating gender differences in body image also revealed higher levels of BID among women compared to men (Borchert & Heinberg, 1996; Purton et al., 2019). Body image misperception (body image distortion) also differs significantly by gender (Chung et al., 2019).

The prevalence of AN and BN in women is overall more known than in men. The general female to male ratio obtained from clinical population for AN and BN is 10/1 (American Psychiatric Association, 2015). Therefore, eating disorders and body image distortion seem to be a fe- male phenomenon. Only body image dissatisfaction appears to concern also men, but probably in a different way than it concerns women.

Men and women produce the same sex hormones, such as estrogens, proges- terone and testosterone, but they differ in their blood concentration, production system, organs, and the apparatus (Lau- retta et al., 2018). Females mainly pro- duce estrogens and progesterone from their ovaries during their cycle pattern and also produce a small amount of tes- tosterone, which is developed by the ovaries and adrenal glands. In contrast, males produce testosterone primarily from their testicles daily. They also pro- duce a small number of estrogens and progesterone, also created by the testi- cles and adrenal glands (Birbaumer & Schmidt, 2010).

The Hypothalamus-Pituitary-Gonadal- Axis (HPG axis) is responsible for the generation of sex hormones, and it con- sists of the Hypothalamus, the Pituitary Gland and the Gonadal Glands. The hy- pothalamus is the highest control station and secretes the hormone GnRH (Gon- adotropin-releasing hormone), which then stimulates the anterior pituitary. As a response to the stimulation, the pitui- tary gland produces LH (luteinizing hor- mone) and FSH (follicle-stimulating hor- mone). Those two stimulate the hor- mone production in the Gonadal Glands, which are the ovaries in females and the testes in males. The female gonads produce the sex hormones estradiol and progesterone, and the male produce testosterone. To regulate all these activ- ities, the HPG axis uses a series of feed- back loops to either stimulate or inhibit the production of GnRH or LH/FSH de- pending on which phase the body is. Because of this feedback mechanism, the hormone concentration across the men- strual cycle has a direct influence on the brain, and, thus, on perception and be- haviour (Birbaumer & Schmidt, 2010; Hill, 2019).

An average menstrual cycle lasts 28 to 29 days and can be divided into four phases. The first menstrual phase, the menstrual phase, starts on the first day of menstruation and ends with the end of menstruation. This phase has a duration between three days and one week. Second phase, the follicular phase, starts at the same time as the menstrual phase. During the first cycle phase, a follicle

matures, resulting in increased estradiol production. The third cycle phase is the ovulation phase and lasts from the 12th day to the 15th day of the menstrual cy- cle. In this phase, the fertilized egg is re- leased from the follicle,

which trans- forms into the corpus luteum and starts producing progesterone. The last cycle phase is the luteal phase, which begins on the 16th and ends on the 28th day of the cycle. During this phase, estradiol and progesterone prepare the uterine mucosa for egg implantation. However, if this does not occur in the third cycle phase, the mucous membrane closes about 14 days after ovulation (Better Health Channel, n.d.; Hill, 2019).

The effects of fluctuations of sex hor- mones during the menstrual cycle on body perception and satisfaction have been investigated in only a few studies, and much research is still needed. Alt- abe & Thompson (1990) tested 60 fe- males aged 17 to 25 on their levels of body image and eating disturbance dur- ing three phases of the menstrual cycle. They found that body image disturbance was higher in the perimenstrual phase (in the menstrual and premenstrual phase compared to the intermenstrual phase) and that women in this cycle phase also overestimated their waist size. The study of Carr-Nangle et al. (1994) showed that BID, measured by the number of body-related negative thoughts, was significantly higher during the perimenstrual phase. Body size per- ception, however, was stable during the different menstrual phases. Jappe & Gardner (2009) found that participants desired a smaller "ideal" body in all three phases. Body size perception did not dif- fer significantly over the phases, in con- trast to BID, which was significantly higher in the premenstrual and in the menstrual phase (compared to the inter- menstrual phase). Teixeira et al. (2013) showed that perceived body size and BID were at their highest during the menstrual phase. There were no significant differences in ideal body size be- tween different menstrual phases. Over- all, participants wished for a smaller size compared to the perceived size. Krohmer et al. (2019) showed that at ov- ulation, naturally cycling women (NC

women) felt more attractive in their bod- ies compared to the late luteal phase. They found an association between the menstrual phase and women's self- rated attractiveness and selective atten- tion when looking at their body parts. So far, there has not been sufficient research regarding the influence of hormo- nal contraception on body image. The primary aim of this study is to analyze whether there is a difference between healthy NC women in the perimenstrual phase, intermenstrual phase, and women on hormonal birth control (HC women) in their body size estimation (BSE) tasks and BID.

Hypothesis

The following hypotheses were pro- posed for the present study:

1. NC women in the intermenstrual phase differ significantly from both NC women in the perimen- strual phase and HC women re- garding both BSE and BID.
2. No significant differences will be found between BSE of NC women in the perimenstrual phase and HC women.
3. No significant differences will be found between BID of NC women in the perimenstrual phase and HC women.

Methods

This study is part of the larger project BODILLUSION, in which the relationship between cardio-visual integration and body image distortion in healthy women is analyzed. Ethics approval was ob- tained by the Ethics Review Panel of the University of Luxembourg on September 24th, 2021 (ERP-20-006). Data were collected at the Clinical Psychophysiol- ogy Laboratory of the Université du Lux- embourg (Campus Belval) between September 2021 and May 2022.

Participants

In total, 35 participants came to the laboratory but several of them had to be excluded based on incorrect or incomplete information they provided in the digital questionnaire regarding their menstrual cycle. Additionally, some participants who did not report menstrual cycles that fall into the normal range of 21 to 35 days (Bull et al., 2019) had to be excluded.

Therefore, the final study sample consisted of 22 healthy women (13 NC women, 9 HC women) aged 18-34 (*M* = 23.36*, SD* = 3.99*).* Eighteen were tested in German and 4 in English. The mean BMI was 21.89 (*SD* = 1.65*)*. The mean duration of the cycle was 28.64 days (*SD* = 2.38), and the mean fluctuation of the cycle was 2.50 days (*SD* = 1.79). The participants were divided into three groups, one group of NC women in their perimenstrual phase (Age: *M*=23.60, *SD*=3.51; BMI: *M*=21.13, *SD*=1.80; Du- ration of cycle: *M*= 28.80*, SD*=3.56; Fluctuation of cycle: *M*=2.80, *SD*=.83) one group of NC women in their inter- menstrual phase (Age: *M*=24.38, *SD*=5.66; BMI: *M*=22.04, *SD*=1.37; Du- ration of cycle: *M*= 29.50*, SD*=2.96; Fluctuation of cycle: *M*=3.25, *SD*=2.05) and one group of HC women (Age: *M*=22.33, *SD*=2.63; BMI: *M*=22.16, *SD*=1.84; Duration of cycle: *M*= 28.00*, SD*=0; Fluctuation of cycle: *M*=1.67, *SD*=1.73). The following data were col- lected to determine how participants were divided into those groups: the last start date of the menstruation and the cycle duration for two menstrual cycles and the hormonal contraception method. For the level of the BMI there was no significant effect, $F(2,19) = .666$, $p = .526$, $\eta_p^2 = .065$. The participants didn't differ significantly also in terms of age (Welch's test revealed $F(2,19) = .540$, $p = .592$, $\eta_p^2 = .054$), as well as regarding mean fluctuation of cycle ($F(2,19) = 1.890$, $p = .178$, $\eta_p^2 = .166$).

Participants were recruited through online advertisements as well as through the distribution of flyers. Volunteers were sent an information booklet with further information and instructions concerning study participation and received a link to the digital questionnaire,

which assessed the in- and exclusion criteria and trait questionnaires. Participants were recruited through online advertisements as well as through the distribution of flyers. Volunteers were sent an information booklet with further information and instructions concerning study participation and received a link to the digital questionnaire, which assessed the in- and exclusion criteria and trait questionnaires.

To participate in the study, participants had to be naturally cycling or on hormonal birth control and speak German or English. If participants had a history of eating disorders (e.g., anorexia nervosa, bulimia nervosa, binge-eating disorder) and/or body dysmorphic disorder (BDD), a history of psychotic disorders, trauma (PTSD), bipolar disorder, substance use disorder, claustrophobia, ADHD or nee- dle phobia, a menstrual disorder, or chronic physical diseases (e.g., epilepsy, diabetes, unmedicated hypo/hyperthy- roidism, cardiac disorder, pacemaker, uncorrected vision, colour blindness, ...), they were excluded. Participants were also excluded if they had a current preg- nancy, current mental disorder or were currently breastfeeding, if their biological sex was male or if their gender was male or other.

If an exclusion criterion was met, partic- ipants of the digital questionnaire were informed that they were not eligible for study participation in the experimental study, thus protecting them from unnec- essary efforts.

To assure that the participants did not have any history of or current eating dis- order diagnosis, the sections for BDD and eating disorders of the Structured Clinical Interview for DSM-IV (SCID-IV, First & Gibbon, 2004) were implemented at the beginning of the laboratory session.

Participants did not eat and only drank still water for the 2 hours before the la- boratory testing session to ensure that they presented with an equal satiation status.

Materials

METRIC BSE TASK. A metric BSE task was implemented as described in a study conducted by Keizer et al. (2016). Participants estimated the circumfer- ence of their shoulders, abdomen, and hips using a piece of string that was placed on the desk so that it would fit ex- actly around the respective body part (e.g., “Please place a piece of string on the table so that the string would fit ex- actly around your *BODY PART*. Please cut the string exactly where it

closes the circle"). The width of the shoulders, ab- domen and hips was estimated by plac- ing two magnetic arrows on a magnetic board rep- resenting the left and right sides of the body (e.g., "Please place these two magnets on the board so that your *BODY PART* fits exactly be- tween the arrows. Your size estimation should represent how you experience your body size"). The order of body parts (shoulders, abdomen, hips) and type of estimation (width, circumfer- ence) was counterbalanced across participants with a balanced Latin square design. The per- centage of misestimation (i.e., either over-or underestimation) for each body part of interest was calculated with the following formula:

$$Body\ Perception\ Index\ (BPI) = \frac{(estimated\ size - actual\ size)}{actual\ size} \times 100$$

DEPICTIVE BSE TASK. Additionally, a de- pic- tive BSE task in virtual reality (BSE VR) (method of adjustment) was imple- mented. For this, a 3D body scan of the participants was taken in the Vitronic VI- TUS 3D body scanner. The 3D body scanner works by laser triangula- tion, which means it projects a light onto the person, and the distance between the sensors and the person is measured. Additionally, a camera captures the per- son's colour infor- mation so that a col- oured 3D figure can be created. The la- ser is entirely harmless to par- ticipants' health, making this a non-invasive pro- cedure. The participants were scanned in A-pose while wearing standardised white leg- gings and a white sports bra provided by the researcher.

Figure 1 Creating a 3D image in stand- ardised clothing.

The 3D body scan of the participants was then imported into a VR environ- ment. The VR en- vironment was pro- grammed in Unity3D (www.unity3d.com). The VR scene consisted of an empty, dark room with a full-length mirror. The HTC Vive Pro Eye virtual reality system was used in this study. In VR, partici- pants could see their 3D body scan in the mirror from their feet to their shoul- ders, leaving out the face so that partici- pants were not distracted by looking at their faces. In a first step, partici- pants could see their 3D body scan in VR for a duration of 90 seconds in order to get ac- quainted with the new environment. Af- ter this, participants filled out a short questionnaire out- side the VR, indicating how positive or negative they experi- enced the 3D body and their arousal level. Following this, the participants proceeded with the BSE task in VR. In this task, participants' distorted 3D body scan was repre- sented in a grey texture. It was shown that BSE tasks in VR pre- sented in a grey texture are more neutral and thus, participants are less in- flu- enced by colour and shadows when es- ti- mating their 3D-Scan (Thaler et al., 2019). The 3D body scan was presented four times in total, twice with a distortion of -20%, and twice with a distortion of +20% of participants' real body size, pre- sented in random order. Participants had to adjust the distorted body with the VR trackers to their perceived body size.

There was no time-limit for participants to work on their BSE. The perceived body size is indicated by the mean per- centage of misestimation across all four trials (to cancel out order effects) rang- ing from -0.8 to 0.8. A percentage of mis- estimation of 0 refers to an accurate body perception.

EATING DISORDER INVENTORY II (EDI- II) (Segura-García et al., 2015). The EDI-II was used as a trait measure of eating disorder symptomatology. The questionnaire consists of 8 subscales, but for the current study only three of them were implemented:

1. Drive for thinness (From test manual: „Strong desire to be thinner, or fear of being fat". The items on this subscale concern intense preoccupa- tion with dieting, mental fixa- tion on weight and fear of gaining weight).
2. Bulimia („A tendency to be preoccupied with binge-eat- ing on a mental as well as on an action level. ")
3. BID (This subscale measures a general BID and the dissat- isfaction with particular body sites which are of the greatest importance for people with eating disorders (waist, hips, thighs). Although BID is very common among young women in Western industrial- ized countries, in its extreme form it is a central issue for patients with anorexia and bulimia.)

Self-Assessment Manikin Scale (SAM- Rating Scale) (Bradley & Lang, 1994). The scale consists of different subscales.

1. Valence. The first subscale measures how positive or nega- tive one feels when seeing the body in 3D. The items of this scale consist of figures with dissatisfied to satisfied faces scaled from 1 to 9 (1 = very positive, 9 = very negative)
2. Arousal. The second subscale assesses how intensely partici- pants feel their arousal when looking at the 3D scan. This scale also consists of the num- bers from 1 to 9 (1 being really intense and 9 not intense).

STATE BODY SATISFACTION. To measure state body satisfaction after looking at the 3D body, three questions were de- veloped by the researchers:

1. *How satisfied are you with your body shape right now?* The scale of the question is the percentage from 0 to 100 in steps of 10% (0% = "*not at all*", 100% = "*completely*".
2. *How satisfied are you with your weight right now?* The scale ranges from 0 to 100 as in question 1, also, here, 0% is being considered as "*not at all*" and 100% is considered as "*completely*".
3. *How is your body feeling right now?*

   The scale ranges from 0 to 100 as in question 1, but with very thin and very thick: 0% is being considered *"very thin"* and 100% *"very thick"*

Procedure of the laboratory session

In the laboratory, participants had to fill out the informed consent form first. Fol- lowing this, the researcher proceeded with the SCID-interview, as well as some questions on hormonal status: They were asked to indicate their last day of menstruation and the duration of their cycle. Then, participants changed into the standardized clothing and the 3D body scan was taken (described above). While one researcher applied the elec- trodes for psychophysiological testing (i.e., ECG, EEG), the second researcher uploaded the 3D body scan into VR and adjusted the virtual room in a way that participants could not see their head in the virtual mirror. After everything was prepared, participants started with the free viewing of their non-distorted body scan in VR. Next, participants filled out the SAM-ratings. Thereafter, participants were asked to do the depictive BSE task in VR, followed by the metric BSE. Finally, they had to answer the State Body Satisfaction questions, fol- lowed by additional tasks that were done in the scope of the larger project

before the session was finished. The re- searcher made sure that the participant was fine, and participants received reim- bursement of 40€ Sodexo gift vouchers for their participation.

Results

To investigate reliability and internal consistency of the different EDI-II scales, Cronbach's alpha was conducted. The drive for thinness subscale consisted of 7 items (α=.645, for the English version α=.863, for the German version α=.612), the bulimia subscale consisted of 7 items (α=.588, for the English version α=.519, for the German version α=.601) and the BID subscale consisted of 9 items (α=.699, for the English version α=.800, for the German version α=.697). This shows that all of the 3 sub- scales have a low reliability. Furthermore, Cronbach's alpha was also con- ducted for the SAM-Rating scale. The subscale valence consisted of 2 items (α=.800, for English version α=.000, for the German version α=.830). The sub- scale arousal consisted of 2 items (α=.754, for English version .640, for German version α=.831). These results show that the subscale valence has a good reliability, and the subscale arousal has an acceptable reliability. The descriptive statistics of the whole sample and for the different groups for BPI can be found in figure 1.

The descriptive statistics of the whole sample and for the different groups for BID, SAM-Rating scores, EDI-II scores can be found in table 1, 2 and 3. The re- sults of the BPI show that overall, women overestimate the width of their body as well as the circumference of their body. It is interesting that NC women in the intermenstrual phase overestimated their body size more than NC women in peri- menstrual phase (same as in the VR- BSE), although in this task women in the perimenstrual phase overestimate and in VR they underestimate.

HC women are the most accurate out of the three groups regarding width estima- tion, but the least accurate in circumfer- ence estimation.

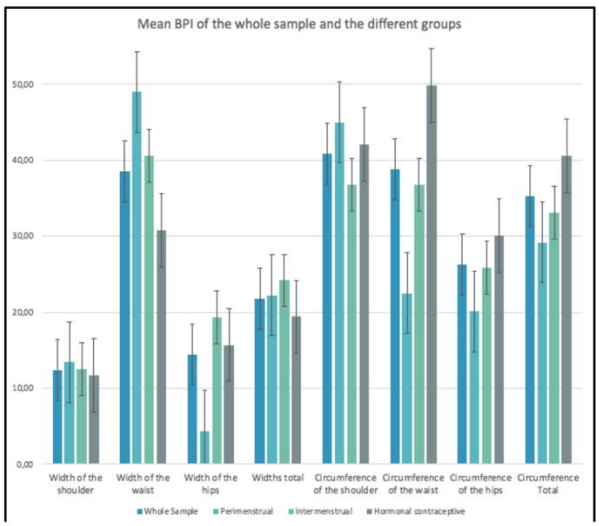


Figure 1 Mean and standard deviation of BPI of the whole sample and the different groups

Table 1: Descriptive statistics of the BID of the whole sample and the different groups in percentage (%)

	Satisfaction with body shape		Satisfaction with body weight		How thin/thick does the body feel?	
	M in %	*SD*	*M* in %	*SD*	*M* in %	*SD*
WS	74.55	17.92	81.82	17.08	57.27	10.32
PM	68	19.24	86	19.49	58	8.37
IM	70	21.28	75	21.38	60	10.69
HC	82.22	12.02	85.56	10.14	54.44	11.30

WS= whole sample, PM= Perimenstrual, IM= Intermenstrual, HC= Hormonal contraception

Table 2: Descriptive Statistics of the SAM-Rating scores for the whole sam- ple and the different groups

Groups	*3D Valence*		*3D Intensity*		*Mood valence*		*Mood intensity*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
WS	2.55	1.13	5.00	1.80	2.14	2.18	5.50	1.97
PM	2.80	1.48	5.00	2.12	2.80	2.05	5.20	1.92
IM	2.88	1.46	5.00	2.20	2.13	.84	6.00	2.27
HC	2.11	1.05	5.00	1.41	1.78	1.09	5.22	1.89

WS= whole sample, PM= Perimenstrual, IM= Intermenstrual, HC= Hormonal contraception

Table 3: Descriptive statistics of the EDI- II scores of the whole sample and the dif- ferent groups

	Drive for thinness		*Bulimia*		*BID*	
	M	*SD*	*M*	*SD*	*M*	*SD*
WS	11.45	3.26	8.55	1.75	21.59	5.81
PM	13.80	3.42	9.00	2.55	29.80	3.83
IM	22.00	3.74	8.13	1.13	26.75	4.53
HC	10.33	1.87	8.67	1.87	20.00	4.33

WS= whole sample, PM= Perimenstrual, IM= Intermenstrual, HC= Hormonal contraception

Table 4: Mean percentage of BSE VR of the sample and the different groups

	M in %	*SD*
Whole Sample	99.46	6.18
Perimenstrual	93.55	4.87
Intermenstrual	105.08	4.63
HC	97.73	3.59

The results in table 4 show that women in their intermenstrual phase overesti- mate their body size by 5.08%, while women in the perimenstrual phase and women on hormonal contraception un- derestimate their body size by 6.45%, and 2.27% respectively.

The correlation coefficient in table 6 in- dicated an understandable trend: the more dissatisfied participants were with their own body, the more negative they experienced their body. No significant correlations were found between the results of SAM Arousal and EDI-Drive for thinness, EDI-BID and BID mean. No significant correlations were found be- tween the results of the BSE task in VR and the metric BSE task.

In order to test the hypotheses and to in- vestigate possible group differences, an ANOVA with one between-subjects fac- tor (hormonal status: NC perimenstrual, NC intermenstrual, HC) was applied. It enabled to identify whether there is a significant difference between means of the considered groups. To find out where specific differences lie, a post hoc test (Scheffe's test) was used. To re- duce the inflated probability of finding a significant result by conducting multiple tests, Bonferroni correction was applied (by α = .05 for 3 tests, corrected α-level would be .0167).

Figure 2 shows that the main effect of BID (calculated as an average index sat- isfaction with 3 considered aspects: body shape, body weight, and general body perception at the moment) was not significant.

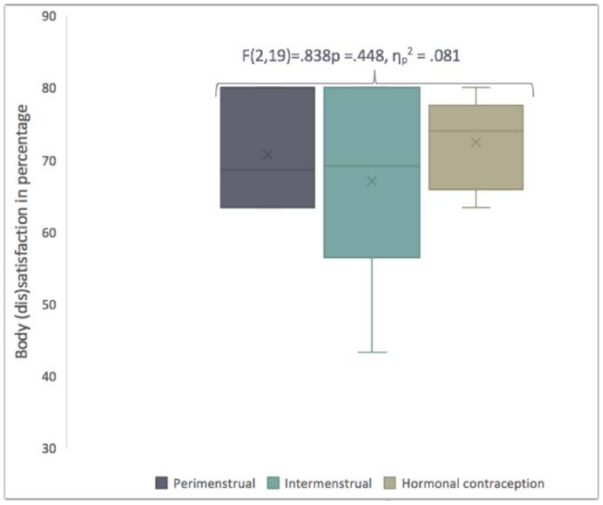


Figure 2 Mean percentage of BID and main effect of NC women in the perimenstrual phase, NC women in their intermenstrual phase and HC women

The correlation coefficient in table 5 in- dicated an understandable trend: the more dissatisfied participants were with their own body, the more negative they experienced their body. No significant correlations were found between the results of SAM Arousal and EDI-Drive for thinness, EDI-BID and BID mean. No significant correlations were found be- tween the results of the BSE task in VR and the metric BSE task. However, BSE VR and BPI circumference of shoulder correlated significantly (c.f. table 7).

In order to test the hypotheses and to in- vestigate possible group differences, an ANOVA with one between-subjects fac- tor (hormonal status: NC perimenstrual, NC intermenstrual, HC) was applied. It enabled to identify whether there is a significant difference between means of the considered groups. To find out where specific differences lie, a post hoc test (Scheffe's test) was used. To re- duce the inflated probability of finding a significant result by conducting multiple tests, Bonferroni correction was applied (by α = .05 for 3 tests, corrected α-level would be .0167).

Table 5: Correlations between EDI-II Drive for thinness, BID, BID mean, Sam Mean Valence and BPI-Circumference of shoulder

	EDI – Drive for thinness	*EDI – BID*	*SAM - Mean Valence*	*BPI- Circumference of shoulder*
	r	*r*	*r*	*r*
EDI – Drive for thinness	1	.681**	.521*	.437*
EDI – BID	.681**	1	.643**	.083
BID Mean	-.740**	.826**	-.572**	-.286

** Correlation is significant at the .01 level
* Correlation is significant at the .05 level

Table 6: Correlations between Mean BSE VR, SAM-Valence, BPI width of shoulder, waist and hips and BPI cir- cumference of shoulder, waist and hips

	Mean BSE VR	*SAM - Mean Va- lence*	*BPI - Width of shoulder*	*BPI- Cir- cumfer- ence of waist*
	r	*r*	*r*	*r*
BPI Width of Waist	-.309	.189	.453*	.040
BPI Width of Hips	.102	.054	.538**	.377
BPI- Cir- cumfer- ence of shoulder	-.424*	.462*	.200	.322
BPI-Cir- cumfer- ence of hips	.030	.255	.136	.633*

** Correlation is significant at the .01 level
* Correlation is significant at the .05 level

Figure 3 indicates that the effect of BSE in VR was significant overall. The post hoc test (Scheffe) revealed that there was a significant difference between the mean BSE_VR scores of participants (NC women) in the intermenstrual phase versus NC women in the perimenstrual phase or HC women. More specifically, the mean BSE_VR score of NC women in the intermenstrual phase (M = 105.08, S = 4.63) was significantly greater than the mean score for NC women in the perimenstrual phase (M = 93.55, SD = 4.87, p = .001) or users of hormonal con- traception (M = 97.73, SD = 3.59, p =.008). However, there was no signifi- cant difference between the mean BSE_VR scores of participants on hor- monal contraception and NC women in the perimenstrual phase (p =.241).

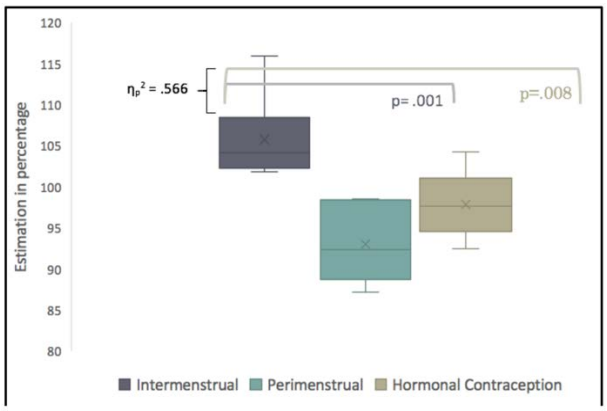


Figure 3 Mean scores of BSE VR and significant levels

of NC women in the intermenstrual, NC women in the perimenstrual phase and HC women

The metric BSE task didn't reveal any signifi- cant main effect (BPI_circumfer- ence: $F(2,19) = 1.075$, $p = .361$, $\eta_p^2 = .102$; BPI_width: $F(2,19) = .185$, $p = .832$, $\eta_p^2 = .019$).

The main effect of SAM-Ratings was not signif- icant overall, too, (Mean Valence: $F(2,19)=.957$, $p=.402$, $\eta_p^2=.091$; Mean Arousal: $F(2,19)=.119$, $p=.888$, $\eta_p^2=.012$. No significant main effect could be found regarding EDI-II (Drive for thinness: $F(2,19) = 3.109$, $p = .068$, $\eta_p^2= .247$; bulimia: $F(2,19) = .390$, $p = .683$, $\eta_p^2 = .039$; BID, $F(2,19) = 1.954$, $p = .169$, $\eta_{p2} = .171$).

Discussion

This study sought to investigate whether there are differences in body image dis- tortion and BID between NC women in different cycle phases and HC women. Significant differences in BSE were found between the different groups, however, only in one of the BSE tasks (BSE-VR), and no differences in BID. Hypothe- sis 1 can, thus, be partly con- firmed. Hypothe- sis 2 can be fully con- firmed, since no group differences could be observed between NC women in the perimenstrual phase and HC women. Hypothesis 3 can also be fully con- firmed. These results show that hormone levels seem not to have an influence on BID, but they seem to have an influence on BSE, but only on estimation from an al- locentric reference frame, while no group differences have been found for the BSE task from an egocentric re- fer- ence frame. Almost all the previous studies found group differences in BID between differ- ent cycle phases (Altabe

& Thompson, 1990; Carr-Nangle et al., 1994; Jappe & Gardner, 2009; Teixeira et al., 2013). Krohmer et al. (2019), how- ever, found no sig- nificant differences in the gaze pattern for the attractive body parts at ovulation and late luteal phase, which is in line with the findings of the present study that BID didn't differ signif- icantly between the different cycle phases. Neverthe- less, they did find sig- nificant differences in the gaze pattern for the unattractive body parts and sig- nificant differences in the self-perceived at- tractiveness. They explained this devi- ation of the findings on attractive body parts compared to unattractive body parts and self-reported at- tractiveness with the assumption "that a bal- anced gaze distribution toward one's own body reflects body satisfaction on a behav- ioral level" (Krohmer et al., 2019), which, unfortu- nately, does not explain the re- sults of the pre- sent study because it does not support the pre- sent findings that there were no significant differences in the self-reported BID (it is only a pos- sible explanation for how the gaze pat- terns fit in with the differences in body satisfac- tion). However, a possible expla- nation could be that the questions as- sessing BID in this study were created by the researchers without previous psy- chometric validation.

The findings of the present study, that there were only small or not significant differences re- garding State Body Satis- faction and the SAM scores, are, thus, a little surprising. The State Body Satisfac- tion questions were shown to partici- pants after the free viewing task and the BSE-VR task, so one could suggest that seeing one's own body in VR might have a positive im- pact on body satisfaction. Mirror exposure is a common therapeu- tic intervention to target BID in women with eating disorders (Griffen et al., 2018), so seeing the 3D scan in VR might have an interventional character already.

Surprisingly, the findings showed that NC women in the perimenstrual phase and HC women underestimated their body, whereas NC women in the inter- menstrual phase over- estimated their body.

Interestingly, both BSE tasks showed different results regarding body image distortion: while participants showed an important overestima- tion for all body parts and all types of estimation (width, circumference) in the metric BSE, par- tic- ipants were, in general, quite accurate in the VR-BSE with a small underestima- tion. Ac- cordingly, there were no signifi- cant correla- tions between the results of the BSE task in VR

and the metric BSE task. This means that these two tasks probably do not measure the same phe- nomenon, which could be shown in previous studies, and meta-analyses (e.g., Mölbert et al., 2017) with the type of es- timation method (metric vs. depictive task) being an important factor. The dif- ferent tasks might involve different pro- cessing steps (explicit vs. implicit; allo- centric vs. egocentric) and therefore do not measure the same subject. BSE might depend on the reference frame of the participant (Monthuy-Blanc et al., 2020)

According to the expectation, the inter- menstrual group differed from the peri- menstrual and hormonal birth control group in the BSE task in VR because the perimenstrual group and hormonal birth control group probably showed similar hormone levels, compared to intermen- strual group, which was the only group with high estrogen. It makes sense that the hormonal contraception group is the "most accurate" because the HC women experience a stable hormonal profile all the time, and thus, should have a "stable" BSE. Nevertheless, it is surprising that the intermenstrual group overestimated their body in the VR_BSE task, while the two other groups showed an underestimation. Research has shown that women in the follicular/intermenstrual phase feel more attractive (Krohmer et al., 2019) and attractiveness is related to skinnier body sizes/ smaller BMI (Cross- ley et al., 2012). However, it is possible that women in the intermenstrual phase overestimate their body in the BSE task in VR, because the vision of their body changes across the menstrual cycle, which might explain why there was are a significant groups differences in the BSE task in VR (visual, "depictive", allocen-

tric), but not the metric BSE task, as vis- ual measurements are dynamic and can naturally fluctuate between one meas- urement and another (de Figueiredo et al., 2021).

In contrast to the BSE task in VR, no group differences could be found in the metric BSE task, which might be the case because of the large variance in scores, which makes it more difficult to find significant group differences.

Strength and Limitations

A strength of this study was to use two different BSE tasks, one metric and one VR. By using these two BSE tasks,, it was possible to directly compare the al- locentric view on the own body and the egocentric view.

Another strength was to use multiple body satisfaction measures. With this method, it can be better ensured that the results don't depend on the measures. When using more than one measure, the risk of response bias can probably be reduced.

The current study had several limitations. The first one being, that the sample size was rather small, resulting in a weak sta- tistical power to perform the analyses: The sample size should have been of N=68 to have a statistical power of 0.2 and a medium effect size (f=0.25). Further, participants only came to the labor- atory once; therefore, only transversal comparisons and not longitudinal com- parisons could be made. Lastly, partici- pants were divided into groups based on their reported menstruation dates and cycle durations. Since no hormonal test- ing was implemented, this group formation could be biased and inaccurate. Future studies should consider letting women come to the laboratory more than once, so it would be possible to make longitudinal comparisons (repeated measures design) and include ovulation tests to determine menstrual cycle phases. Furthermore, the hormo- nal contraception group should be re- stricted to one hormonal contraception method (e.g., combination pill). It would also be of interest to include more con-

trol variables regarding emotions and af- fect during the laboratory session to ac- count for mood differences. Finally, fu- ture studies should consider comparing visual body processing (with EEG: event related potentials, N170) in both refer- ence frames and analyse the influence of menstrual cycle phase on the latter. Finally, when looking at the results, researchers discovered that a relatively large number of women lack body liter- acy regarding their menstrual cycle. Quite a few of them didn't

know their cy- cle lengths, or did confuse it with the du- ration of their menses. Even some women who were on hormonal birth con- trol (e.g., birth control pills) weren't in- formed about what their cycle length is. In the future, it would be necessary to better educate women and young girls about the menstrual cycle, p. e. in school programs.

To sum up, there are not many studies investigating the female cycle and its in- fluence on women's well-being and sat- isfaction with their bodies. The few ones that exist are already relatively old and lack a comprehensive research design. The present study sought to investigate whether the menstrual cycle, as well as hormonal contraception, have an influ- ence on body image distortion and BID in healthy women. While finding some interesting results, the evidence remains inconclusive and more detailed research is needed to unravel the effects of hor- monal status on body image. Further research in this domain is of great im- portance because the menstrual cycle and body image disturbance concerns no less than half of the world's population.

References

Administration Substance Abuse and Mental Health Services. (2016, June). *Table 19, DSM-IV to DSM-5 Ano- rexia Nervosa Comparison*.

Altabe, M., & Thompson, J. K. (1990). Menstrual cycle, body image, and eating disturbance. *International Journal of Eating Disorders*, *9*(4). https://doi.org/10.1002/1098 108X(199007)9:4<395::AID-

EAT2260090405>3.0.CO;2-E American Psychiatric Association. (2015). DSM-V: Troubles des con- duites alimentaires et de l'ingestion d'aliments. In Elsevier Masson SAS (Ed.), *DSM-5 Manuel diagnostique et statistique des troubles mentaux* (5e édition, pp. 444–456). APA Dictionary of Psychology. (n.d.). *Body Image*. Retrieved June 12, 2022, from https://dictionary.apa.org/ Better Health Channel. (n.d.). *Menstrual Cycle*. Retrieved June 10, 2022, from https://www.bet- terhealth.vic.gov.au/health/condi- tion- sandtreatments/menstrual-cycle

Birbaumer Niels, & Schmidt Robert F. (2010). Biologische Psychologie: En- dokrine Systeme. In Springer Berlin Heidelberg (Ed.), *Biologische Psy- chologie* (7. Edition, pp. 135–136).

Borchert, J., & Heinberg, L. (1996). Gen- der schema and gender role discrep- ancy as correlates of body image. *Journal of Psychology: Interdiscipli- nary and Applied*, *130*(5), 547–559. https://doi.org/10.1080/00223980.19 96.9915021

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assess- ment manikin and the semantic dif- ferential. *Journal of Behavior Ther- apy and Experimental Psychiatry*, *25*(1), 49–59. https://doi.org/10.1016/0005- 7916(94)90063-9

Bull, J. R., Rowland, S. P., Scherwitzl, E. B., Scherwitzl, R., Danielsson, K. G., & Harper, J. (2019). Real-world men- strual cycle characteristics of more than 600,000 menstrual cycles. *Npj Digital Medicine*, *2*(1). https://doi.org/10.1038/s41746- 019- 0152-7

Carr-Nangle, R. E., Johnson, W. G., Bergeron, K. C., & Nangle, D. W. (1994). Body image changes over the menstrual cycle in normal women. *International Journal of Eating Disor ders*, *16*(3). https://doi.org/10.1002/1098- 108X(199411)16:3<267::AID-EAT2260160307>3.0.CO;2-Y

Crossley, K. L., Cornelissen, P. L., & To- vée, M. J. (2012). What Is an Attrac- tive Body? Using an Interactive 3D Program

to Create the Ideal Body for You and Your Partner. *PLoS ONE*, *7*(11). https://doi.org/10.1371/journal.pone.0050601

De Figueiredo, B. G. D., Rezende, M. T. C., dos Santos, N. A., & de Andrade, M. J. O. (2021). Mapping changes in women's visual functions during the menstrual cycle: Narrative review. In *Sao Paulo Medical Journal* (Vol. 139, Issue 6). https://doi.org/10.1590/1516-3180.2020.0474.R2.03052021

First, M. B., & Gibbon, M. (2004). The Structured Clinical Interview for DSM-IV Axis I Disorders (SCID-I) and the Structured Clinical Interview for DSM-IV Axis II Disorders (SCID- II). In *Comprehensive Handbook of Psychological Assessment, Person- ality Assessment* (Vol. 2).

Frederick, D. A., Peplau, L. A., & Lever, J. (2006). The swimsuit issue: Corre- lates of body image in a sample of 52,677 heterosexual adults. *Body Image*, *3*(4). https://doi.org/10.1016/j.bodyim.2006.08.002

Fuentes, C. T., Longo, M. R., & Haggard, P. (2013). Body image distortions in healthy adults. *Acta Psychologica*, *144*(2). https://doi.org/10.1016/j.actpsy.2013.06.012

Gaudio, S., & Quattrocchi, C. C. (2012). Neural basis of a multidimensional model of body image distortion in an- orexia nervosa. In *Neuroscience and Biobehavioral Reviews* (Vol. 36, Is- sue 8). https://doi.org/10.1016/j.neu- biorev.2012.05.003

Griffen, T. C., Naumann, E., & Hilde- brandt, T. (2018). Mirror exposure therapy for body image disturbances and eating disorders: A review. In *Clinical Psychology Review* (Vol. 65). https://doi.org/10.1016/j.cpr.2018.08.006

Jappe, L. M., & Gardner, R. M. (2009). Body-image perception and dissatis- faction throughout phases of the fe- male menstrual cycle. *Perceptual and Motor Skills*, *108*(1). https://doi.org/10.2466/PMS.108.1.7 4-80

Keizer, A., van Elburg, A., Helms, R., & Dijkerman, H. C. (2016). A virtual re- ality full body illusion improves body image disturbance in anorexia ner- vosa. *PLoS ONE*, *11*(10). https://doi.org/10.1371/journal.pone.0163921

Kring, A. M., Johnson, S. L., & Hau- tzinger, M. (2019). Klinische Psychol- ogie. Ein Lehrbuch. In Beltz (Ed.), *Behaviour Research and Therapy* (9. Auflage, pp. 347–351).

Krohmer, K., Derntl, B., & Svaldi, J. (2019). Hormones matter? Associa- tion of the menstrual cycle with selec- tive attention for liked and disliked body parts. *Frontiers in Psychology*, *10*(APR). https://doi.org/10.3389/fpsyg.2019.00851

Lauretta, R., Sansone, M., Sansone, A., Romanelli, F., & Appetecchia, M. (2018). Gender in endocrine dis- eases: Role of sex gonadal hor- mones. In *International Journal of Endocrinology* (Vol. 2018). https://doi.org/10.1155/2018/484737 6

Mölbert, S. C., Klein, L., Thaler, A., Mohler, B. J., Brozzo, C., Martus, P., Karnath, H. O., Zipfel, S., & Giel, K. E. (2017). Depictive and metric body size estimation in anorexia nervosa and bulimia nervosa: A systematic review and meta-analysis. In *Clinical Psychology Review* (Vol. 57). https://doi.org/10.1016/j.cpr.2017.08.005

Monthuy-Blanc, J., Bouchard, S., Ouel- let, M., Corno, G., Iceta, S., & Rous- seau, M. (2020). "eLoriCorps Immer- sive Body

Rating Scale": Exploring the Assessment of Body Image Dis- turbances from Allo-centric and Ego- centric Perspectives. *Journal of Clin- ical Medicine*, *9*(9). https://doi.org/10.3390/jcm9092926

Purton, T., Mond, J., Cicero, D., Wagner, A., Stefano, E., Rand-Giovannetti, D., & Lat-ner, J. (2019). Body dissatisfac- tion, in-ternalized weight bias and quality of life in young men and women. *Quality of Life Research*, *28*(7), 1825–1833. https://doi.org/10.1007/s11136-01902140-w

Sarah E. Hill, P. (2019). This is Your Brain on Birth Control. In *Penguin Random House LLC New York*. Segura-García, C., Aloi, M., Rania, M., Ciambrone, P., Palmieri, A., Pugliese, V., Ruiz Moruno, A. J., & de Fazio, P. (2015). Ability of EDI-2 and EDI-3 to correctly identify patients and sub-jects at risk for eating disorders. *Eat- ing Behaviors*, *19*. https://doi.org/10.1016/j.eat-beh.2015.06.010

Teixeira, A. L. S., Dias, M. R. C., Dama- sceno, V. O., Lamounier, J. A., & Gardner, R. M. (2013). Association between different phases of men- strual cycle and body im-age measures of perceived size, ideal size, and body dissatisfaction. *Per- cep-tual and Motor Skills*, *117*(3). https://doi.org/10.2466/24.27.PMS.117x31z1

Thaler, A. et al., 2019. The Influence of Visual Perspective on Body Size Es- timation in Immersive Virtual Reality. In Association for Computing Ma- chinery (ACM), pp. 1–12.

The influence of language and emotions on the cognitive performance of children

Silvia Bulzacchi, Margot Ewen, Kim Häfner and Eline Liang

Supervisors: Dr. Andreia Costa and Maïte Franco

This paper analyzes the influence of language and emotions on cognition and more precisely whether there is a difference between the use of the mother tongue or a second language during an emotion-eliciting (frustrating) situation in a subsequent cognitive performance (memory) of children aged between 6 to 12 years old. Given that past research demonstrated that having to regulate emotions can restrict resources for accomplishing other concurrent cognitive tasks (Richards & Gross, 2000; Wentzel & Miele, 2016) and as multilingual people show less emotional reactivity by distancing themselves when their non-dominant language is being used (Lindquist et al., 2013), we expected that children would have worse cognitive performance if their mother tongue is used. The sample (N = 24) was composed of two equal-sized groups. One group consisted of participants having German as their mother tongue (L1 group) while the other group included individuals speaking German as a second language (L2 group). In order to elicit an emotional state, namely frustration in the children, we made use of a disappointment paradigm (adapted from Cole et al., 1994; Saarni, 1984) which was followed up with a memory task to evaluate their cognitive performance and an interview in German regarding their emotion regulation. Results show that participants of the L2 group had a higher emotional reactivity than the L1 group after receiving their least preferred item. In addition, participants in the L1 group had significantly higher IQ scores. Language proficiency (either first or second language) and emotionality during the memory task had, however in contradiction to our hypothesis, no significant impact on cognitive performance. Moreover, the results demonstrated that IQ was the best predictor for the memory task score.

1. Introduction

More than half of the population in Europe is at least bilingual (Grosjean, 2021). Especially countries like Luxembourg, Switzerland and the Netherlands are characterised by their bilingualism and/or multilingualism. In particular in the Luxembourgish education system, one can be confronted with subjects in many languages. Luxembourgish is the main language of instruction in the first cycle of primary school. However, the percentage of students speaking Luxembourgish as their mother tongue counts only 35% (MENJE, 2019, as cited in Ugen et al., 2021). The main other languages spoken at home by pupils in Luxembourg are “Portuguese (23%), French (8%), and South Slavic languages (4%)” (Martini et al., 2021). Moreover, as the literacy instructions in Luxembourg's primary school system continue to be in German, but only circa 2% of students speak German as their main language at home, this can have an impact on the pupils’ performance in school subjects (MENJE, 2019, as cited in Ugen et al., 2021).

Language and cognition

Previous research has shown that in Luxembourg, children speaking Luxembourgish at home and for whom the main instruction language is German show a better performance in main school subjects such as maths and German Reading Comprehension (Martini et al., 2021). This phenomenon is also described by Hoffman et al. (2018), because students speaking Luxembourgish or German as first language display more stable and positive developmental progression in reading competences.
Given the strong migrant landscape of Luxembourg, it is important to understand the potential factors which might influence cognitive performance,

especially children's learning and academic achievement. In fact, there is a close relationship between language and cognition and therefore thoughts are influenced by language as well (Boroditsky et al., 2003).
Bilingual people show better cognitive performance by using executive functioning and executive control, for instance inhibition, for tasks unrelated to language acquisition (Baumgart & Billick, 2018). However, it has to be noted that some of the effects found in the "bilingual advantage" studies are mostly due to the participants' socio-economic status (SES). Morton and Harper (2007) suggest that controlling for differences in ethnicity and SES can attenuate the bilingual advantage in cognitive control as bilingual and monolingual children performed the same in their study when the socio-economic status was identical for both language groups. Furthermore, when the socio-economic status was not alike, then children coming from higher SES families had an advantage over children from lower SES families, which supports this statement.

Emotion and cognition

The same way as language, emotions also are an important factor in terms of performance in school settings. "Emotions are functionally important for students' motivation, academic success and personality development" (Wentzel & Miele, 2016).
Research has demonstrated that emotions have an impact on cognitive performance. Emotional responses to affective stimuli for example can lead to poor performance in goal-directed tasks. In particular, emotional distractors, be it negative or positive stimuli, can disrupt goal-directed processing (Blair et al., 2007). Moreover, emotional states consume cognitive resources and therefore reduce cognitive performance. In other words, focusing attention on emotions leads to restricted availability of resources for concentrating on the task at hand (Ellis & Ashbrook, 1988; Meinhardt & Pekrun, 2003, as cited in Wentzel & Miele, 2016). As mental resources are necessary for self-regulating emotions, performance could as a result be impaired because attention is a limited resource. Taking memory as an example for the effects emotions can have on cognitive performance, research has shown that antecedent-focused and response-focused regulation strategies each have different impacts on cognitive processes. Expressive suppression, a response-focused strategy, for instance, leads to poorer memory for emotional events while reappraisal, an antecedent-focused strategy, does not influence memory performance as the situation's emotional reality is changed beforehand and thus no further cognitive work is necessary (Richards & Gross, 2000). These results demonstrate that students' emotions and their down-regulation could have an impact on remembering what they are taught in school. For instance, reacting to a negative emotion such as frustration with disruptive, interfering responses was postulated to reduce the quality of performance on activities following frustration (Child & Waterhouse, 1953). Fox and Spector (1999) supported the assumption that organisational frustrating events are associated with counterproductive emotion and behavioural responses resulting from affective responses to frustration. As this was only analysed in the occupational setting or in university students, it would be interesting to see if the same impact of frustration on cognitive performance persists when exploring the effect in primary school children.

Language and emotion

Regarding the relation between language and emotion there is great controversy as there are two main opposite assumptions. On the one hand there is the modular point of view believing that language and emotions are two independent constructs as, according to supporters of this theory like Chomsky or Jerry Fodor, the mind is composed of innate neural structures or mental modules which have distinct functions (Grenkowski, 2012).

On the other hand, psychological constructivism states that these processes are connected, and that the spoken language can influence the emotional reactions in a given situation. According to Caldwell-Harris (2014) our intense emotions are often expressed in our native language. In most cases we have a stronger connection to our mother tongue than to a second language. As described in the article of Lindquist et al. (2015), multilingual people can implicitly regulate their emotions by distancing themselves when someone speaks their non-dominant language and therefore show less emotional reactivity. A study by Keysar et al. (2012) suggests that thinking in a different language than your mother tongue may reduce decision biases which would support the thesis that speaking a foreign language provides a greater emotional and cognitive distance than a native tongue.
Moreover, Bond and Lai (1986) revealed that it is for instance easier to speak about embarrassing topics such as embarrassing personal events in a second language than in our mother tongue.
On the other hand, previous research found that bilinguals recalled more emotional meaning words than neutral meaning words in their mother tongue (Ayçiçegi & Harris, 2004). The same effect was observed when using their second language, meaning that regardless of the language used, people recalled more emotional meaning words than neutral ones. These results suggest that emotion words, in particular taboo words, could have an advantage on memory performance in both languages, especially in regard of recall and recognition because such words contain important emotional associations. These findings show that emotions could lead to better cognitive performance concerning memory regardless of people using their mother tongue or a second language.

Most previous research focuses either on the interaction between emotions and cognitive performance or between emotions and language. There is however currently a lack of research combining all these factors (i.e., emotion, language and cognitive performance). Hence, the aim of this project is to understand the influence that emotions, multilingualism and the use of language in dealing with emotions can have on children's cognition. As we previously described German is the main instruction language in primary schools in Luxembourg and only the minority of students possess it as their mother tongue. For this reason, our research project could be very interesting for the school system in Luxembourg, because it would shed light on the question how language proficiency influences cognitive performance in an emotion-eliciting situation and as a result the performance of children in school subjects.

In line with this objective, the following research question was established: “Is there a difference between the use of the mother tongue or a second language during an emotion-eliciting (frustrating) situation in the subsequent cognitive performance (memory) of primary school-aged children?” As previously described, research has shown that there is a connection between language and thought (Boroditsky et al., 2003). Moreover, research has demonstrated that both positive and negative emotional distractors can lead to poor cognitive performance in certain goal-directed tasks (Blair et al., 2007). Emotion regulation consumes cognitive resources and therefore leads to the restricted availability of resources for other tasks (Ellis & Ashbrook, 1988; Meinhardt & Pekrun, 2003).
As in other previously described research (Caldwell-Harris, 2014; Lindquist et al., 2015; Keysar et al., 2012; Bond & Lai, 1986) we hypothesise that children will have a worse cognitive performance if they use their mother tongue during an emotion eliciting situation than if they use a second language as they can distance themselves from the situation if they are using their non-dominant language.

2. Methods

The most important inclusion criteria in order to participate in our study was language proficiency in German. For this reason, we worked with two groups containing an equal number of 12 participants each; one group consisted of participants that have German as their mother tongue and the second group has been learning German as their second language. The participants that learned German as a second language needed to be fluent enough to understand and speak basic German in order to participate. To be able to communicate properly outside of the procedures, the mother tongue of the participant was also spoken by the respective researcher in charge of their assessment and testing. In this case, the languages were as follows: French, English, Luxembourgish and Italian.

2.1 Participants

The sample consisted of 24 participants in total and counted 58,3% boys and 41,7% girls. The mean overall age was 9 (*SD* = 2.03) with an age range from 6 to 12. There were 6 girls and 6 boys in the L1 group (participants having German as a first language) and 4 girls and 8 boys in the L2 group (German as a second language). With 33,3 % of the overall participants, thus 66.6% of the participants in the L2 group of our study having indicated Luxembourgish as their mother tongue, this language dominated in the L2 group.

In addition, another important inclusion criterion was the need for an IQ that did not lie in the below average range for the age of the participant, because the study included an assessment of their cognitive performance. This was controlled for by the researchers having a conversation with the parents about their child and in some cases also by talking to the children beforehand. If the parents did not mention any abnormalities or the child was able to hold a proper conversation, they were deemed suitable for this study. Furthermore, it was required to have normal or corrected-to-normal hearing and vision, as the different procedures required the participants to see, as well as hear the instructions to the different exercises and questions. Moreover, participants who had an official diagnosis of a disorder such as, for example, anxiety, depression, autism, ADHD, or dyslexia were excluded from the sample.

2.2 Materials

Demographic information was assessed via a self-report questionnaire and included age, nationality, household income and education of the parents. Additionally, the participants' date of birth, gender and their substance intake right before the study (e.g., caffeine, chocolate, medicaments) were listed by the parents. This demographic information can however be considered as incomplete, as the children's year of school attendance is missing for most of the participants.

In order to assess children's habitual emotion regulation, the parent form of the Emotion Regulation and Social Skills Questionnaire (ERSSQ-P, Butterworth et al., 2014) was filled out by the parents with a rating scale designed specifically to assess the social skills of young people with Autism Spectrum Disorder (ASD). It has a total of 27 items, which include statements such as the following examples: "Is aware of other people's thoughts and feelings.", or "Recognizes when other people are being sarcastic or teasing". It uses a 5-point Likert scale ranging from 0 (never) to 4 (always), so the possible total score ranges from 0 to 104. As for the interpretation of the score, the higher the score, the higher the child's competence in specific emotion recognition, emotion regulation and social skills. The ERSSQ-P has high internal consistency ($\alpha = .90$). Pearson correlations for concurrent and criterion validity are significant for parent and teacher forms. Specifically, the ERSSQ-P exhibited a strong positive association ($r = .86$, $p < .001$, $n = 61$) (Butterworth et al., 2014).

Regarding the severity of the children's autistic levels, the second edition of the Social Responsiveness Scale (SRS-2, Constantino, 2012) was used. It has 65 items such as the following two examples: "Expressions on his or her face don't match what he or she is saying", or "Is able to communicate his or her feelings to others" and the child can be rated by teachers and parents. The SRS-2 generates raw scores for 5 different domains, which are converted into T-scores. Those can be organised by gender and age and the possible scores range from 32-114. A score below 59 is considered low to no symptomology and the higher the score, the more likely it is for the child to be diagnosed with ASD. In the context of this study, this measure was in particular used to identify possible social communication difficulties in the children. For the full SRS-2 the internal consistency (Cronbach's α = 0.935), and the test-retest reliability (Shrout–Fleiss = 0.944, Winer reliability = 0.944) have been proven to be excellent (Gergoudis et al., 2020).

The children's level of alexithymia was measured by the Alexithymia Questionnaire for Children – Parent (AQC-P, Costa et al., 2017). The questionnaire consists of 20 items such as: "Finds it difficult to say how he/she feels inside." or "He/she often does not know why he/she is angry", which are rated on a 3-point scale from 1 (not true) to 3 (true). The answer "does not apply" is also scored with one point. Scores can range between 20 and 60 with higher scores indicating a higher likelihood for alexithymia. The AQC-P showed acceptable omega reliability values for the total scores (ω = .870) in terms of internal consistency. The total AQC and AQC-P scores were also significantly correlated (r = .325, p < .001) (Brown et al., 2021).

The language proficiency of the participants was indicated by the parents by completing the language history questionnaire (LHQ-3, Li et al., 2020) which is an important tool for assessing the linguistic background and language proficiency of multilinguals or second language learners. Parents first indicated their children's general language learning skills on a scale from 1 (very poor) to 7 (excellent). Secondly, they rated their child's current ability in terms of listening, speaking, reading and writing for all languages the child had learned so far. The validity and reliability of the LHQ questions have been tested as well. As demonstrated in Grant and Li (2019) for example, bilingual participants' verbal fluency scores in Spanish were significantly correlated with their LHQ-based self-rated proficiency scores (p = .039, r = .36) (Li et al., 2020).

Moreover, to test their proficiency level in German, we used four of the ten subtests ("Bildbenennung", "Handlungssequenzen", "Fragen zum Text", "Satzbildung") of the SET 5-10 (Sprachstandserhebungstest für Kinder im Alter zwischen 5 und 10 Jahren, Petermann, 2018). The majority of subtests have good internal consistency with Cronbach's alpha between α = .71 and .91. Analyses of the criterion validity of the procedure revealed medium to high correlations between the subtests of the SET 5-10 and other test procedures that measure comparable constructs. The results speak for the validity of the test. The first subtest consisted of naming and describing several different images that were presented to the participants on cards to determine the extent of their vocabulary. After this, the participants' language comprehension was tested by presenting them with small figures (a woman, a man, a girl, a boy, various animals, a tree, and a bench). The participants had then to use the figures to reenact short sentences, which were read out by the researcher. For the next subtest, short texts were read to the children (four texts for the 6-year-old participants, and five texts for the 7- to 12-year-old participants). Following each story, they were presented with questions and given three alternative answers from which they had to choose the one that best matched the text. For the last task of this test, they needed to form grammatically correct and reasonable sentences from words that were read out

loud by the researcher. This task required knowledge of morphology, syntax, and a sufficiently large vocabulary. The participants' answers to the first and last subtests were recorded so they could be listened to again at the time of the evaluation, as to avoid any possible misunderstandings on the researcher's part. The recordings were deleted after successful transcription of the participants' responses.

To assess the participants' cognitive performance, we used the short version of the WNV test (Wechsler Nonverbal Scale of Ability, Wechsler & Naglieri, 2006). The reliability of the subtests ranges from *r* = .72 (arranging pictures) to *r* = .90 (matrices test). Reliability for the 2-subtest battery is .90. Retest reliability ranges from .78 (4;0 to 7;11) to .89 (8;0 to 21;11). We applied the following subtests: Matrices Test, shape recognition, and visual-spatial memory span. Participants of all age groups completed the Matrices Test. In addition, participants under 8 years old did the recognition task, while those who were 8 years or older completed the visual-spatial memory task instead. For the Matrices subtest, the participants are confronted with pictures of colourful geometrical figures that follow a pattern but are missing one element. The participants were asked to point to the option of the different figures shown that would best complete the pattern. In the Recognition subtest the participants had to look at an image of a shape for 3 seconds. After this, the children had to choose the same shape they had seen beforehand from different options. For the visual-spatial memory span, the participants had to touch different blocks on a board, standing between them and the researcher, in the same and reversed order demonstrated by the examiner. The researcher showed the participants images with the instructions for each WNV task in order to nonverbally communicate what they would have to do. If the participants had problems understanding the instructions, the researcher was allowed to additionally verbally explain the task at hand.

In order to elicit frustration in the children, we made use of a disappointment paradigm (adapted from Cole et al., 1994; Saarni, 1984). First, the participants were asked to rate their preference for different objects, which included sweets, play dough, a pop-it fidget toy, juice, and a slightly burned down candle from 1 (most preferred) to 5 (least preferred). Then they were asked to do a 5-piece puzzle of a car with the knowledge that they would get their most preferred item if they would do very well in solving the puzzle. After having finished the puzzle, all participants, independently of their performance, received their least preferred item, which should frustrate them and elicit a negative disposition.

To assess the participants' emotional state at baseline, namely before the frustration task and after the emotion-eliciting situation, the rating for affective dimensions valence scale (1 = very sad to 5 = very happy) of the Self-Assessment Manikin (Bradley & Lang, 1994) and an interview regarding the participants' emotion regulation were used. The SAM consists of graphic characters with different facial expressions including a smiling, happy figure, a less happy figure, a neutral figure, a not so happy one and a figure that is not happy at all.
The original article describing the SAM reports good psychometric properties (Bradley & Lang, 1994). A very high correlation was found between the SAM items and those of other verbal-based measurement instruments, including high reliability across age (Graziotin et al., 2015).

Additionally, we assessed the children's heart rate with a wrist-worn heart rate monitor (i.e., the Apple® watch series 6) to control for physiological indications of changes in the participants' emotional state. The measurement started before the disappointment paradigm with a five minute resting time to assess a baseline and continued until the end of the study. To be more precise, we had a look at the exact heart rate at the following time

points: At the beginning and at the end of the five minute baseline measurement, after the participants finished the puzzle task, when they received their least preferred item and when the emotion interview began.

To evaluate the participants' cognitive performance (i.e. memory for visual content), the children watched a short non-verbal educational video of 6 minutes and 14 seconds (Wasser in der Wüste - Die Sendung mit der Maus® by the Westdeutscher Rundfunk Köln) and were subsequently asked 14 non-verbal questions about the video (e.g. to recall certain details about the video; to sort different images directly taken from the video into the correct order; to identify which image did not appear in the video). The video questionnaire was created specifically for this study by our research group, and thus was not a validated measure.

In order to measure the effect of language during the emotion-eliciting situation, the researcher conducted an interview with the participants in which they were asked to express how they felt during the disappointment paradigm and about their emotion regulation. The 12 questions (e.g., "When you received your least preferred item, did you do something to make yourself feel better?"; "Did you try to hide your disappointment?"; "When you are sad, what do you usually do to make yourself feel better?") were formulated by our research group and were not validated. The interview was held in German for all the participants. Due to the separation of our study sample into two groups, this interview, which was recorded, was in their mother tongue for the L1 group and in their second language for the L2 group.

2.3 Procedure

The testing phase for the study took place during a period of two weeks in the premises of the Maison des Sciences Humaines building on the Belval Campus (PAT-LAB) of the University of Luxemburg or at the participants' request, at a safe and undisturbed location closer to them. We always made sure that the environment in which the testing took place was quiet to give the participants the possibility to concentrate properly on the tasks on hand. All the procedures were conducted in person and in the presence of an examiner. After we found children who met our inclusion criteria, we contacted the parents for more information about their child and to set a date for the testing. The appointment consisted of one session lasting up to one hour for each participant.

Before the actual experiment began, the participants' parents had received some information about the content and objectives of the study and any further questions were answered. The information sheet explained that the participation would be anonymous and that the participants would be confronted with an emotion-eliciting situation for a short period of time. The pseudonymity of the test subjects was given, by the design of the study, as each subject received an individual and randomly generated identification number with the help of the software Excel. After having agreed that their children could take part in the experiment and signing the legal documents, the researcher started with the tests. It should be noted that the experiment could have been stopped at any moment, and it was possible to ask for their data to be deleted even after the study had been conducted with no negative consequences for neither the participants nor their parents.

Before every testing, the parents and the participants were asked to sign the consent form. Additionally, one of the parents filled out the questionnaires regarding demographics, autistic symptomatology (SRS-2, Constantino, 2012), alexithymia (AQC-P, Costa et al., 2017), and emotion regulation (ERSSQ, Butterworth et al., 2004) of their children. Furthermore, the parents completed a part of the language history questionnaire (LHQ-3, Li et al., 2020). This was either done at home or on-site and took approximately 35 minutes for each child.

To begin the test session, the participants started with the different WNV tasks. After finishing the subtests of the WNV test, the participants continued with the four tasks of the SET 5-10. For the SET5-10 exercises, the researcher as well as the participant communicated with each other in German. As the answers of the participants were recorded for this part, the children were given a short explanation as to why the researcher was recording their voice and were asked to start and stop the recording on their own in order to make it more comfortable for them. The next step consisted of the participants putting on an apple watch. They then had to sit as calmly as possible for a duration of 5 minutes to assess a baseline of their heart rate. The measurement of their heart rate continued throughout the rest of the experiment. Moreover, the participants were asked to indicate how they were feeling after the 5 minutes had passed by using the Self-Assessment Manikin. Then, in order to elicit frustration in the children the disappointment paradigm was used. After initiating frustration in them, they were asked again how they were feeling after receiving their least preferred object. Next, the participants had to watch a short educational video and answer some memory questions about it afterwards. This task was then followed by an interview, where the researcher asked the participants about their emotion regulation during the experiment and in general.

At the end of the study, the children got an explanation as to why they did not get their preferred item even though they did well in solving the puzzle, to make sure that the elicited state of frustration would only be temporary. Finally, the experiment ended with the participants receiving all the objects they had to rank before solving the puzzle, except for the candle, as compensation for their participation in the study and they were told that they had performed well.

2.4 Statistical analysis

All collected data, including the parent's answers to the different questionnaires, the scores reached in the WNV intelligence test and in the SET 5-10, and the questionnaire regarding the video, as well as the heart data and the children's answers to the interview, were collected in a data set in the statistics program "IBM® SPSS Statistics". First, we conducted descriptive analysis such as the mean, standard deviation, number of participants and frequencies to get a general idea of the study sample. Furthermore, different statistical tests were conducted to compare the means of the two groups. The comparison of means included the results of the different tests that we conducted and the information we had from the parents about their children. A regression analysis was used to see whether emotionality, the WNV total score and the group affiliation have an impact on cognitive performance.

3. Results

We analysed the cognitive performance of 24 children after they've just experienced an emotion-eliciting situation to see whether they show a worse performance depending on whether the language used is their first or second language. The comparison was between two groups, one with children that have German as their first language (n = 12) and the other group with children that have German as their second language (n = 12).

Table 1: Age distribution, gender distribution and socio-economic status in the two groups.

	L1	L2	Significance
Age	M = 8.46, SD = 1.79	M = 9.54, SD = 2.19	$t(22)$ = -1.32, p = .199
Gender distribution	6 girls, 6 boys	4 girls, 8 boys	$X^2(1, N=24)$ = .69 p = .408
Income levels of the parents	M = 5.27, SD = 1.01	M = 5.25, SD = 1.14	$t(21)$ = .05 p = .96
Education level of the parents	MR = 17.33	MR = 7.67	U = 14.00, z = -3.49 p = .00

WNV score	*M* = 116.5, *SD* = 13.81	*M* = 99.91, *SD* = 20.70	*t*(22) = 2.31, *p* = .031
SET5-10 score	*M* = 61.85, *SD* = 11.74	*M* = 50.50, *SD* = 16.37	*t*(22) = 1.95, *p* = .06

As indicated in table 1, no significant differences between the two groups were found concerning age and gender distribution. A Student's *t*-test was conducted to compare the age of the participants and a Chi-Square test of independence was conducted to compare the gender distribution.

Furthermore, the results of a Student's *t*-test revealed no significant differences between the income levels of the parents from the children of the two groups. A Mann Whitney test to compare the highest degree of the legal guardians was conducted to compare the second indicator for the socio-economic status of the families of our participants. For each child, the average of the education level of both legal guardians was calculated. Education levels ranked from 1 (Fundamental/primary education) to 6 (master's degree). The results show significantly higher education levels of the parents of children in group L1. However, both groups show a similarly high income with at least half of the families having chosen the highest possible option on our scale. In the L1 group, 7 parents indicated the highest possible score and in the L2 group, 6 parents indicated the highest possible score. Broader differences can be observed regarding the highest degree. Data about the highest degree of the parents (*N* = 48), with 24 in each of the groups, were collected. More than 80 % of parents in group L1 specified that they have obtained a master's degree, which was only the case for one third of the parents in group L2.

In addition, the scores reached in the WNV intelligence scale were higher in the L1 group. An independent sample t-test showed that this difference is significant.

Results of the SET5-10 scores to measure the German language proficiency showed a marginally significant difference between the two groups with statistically higher results of children with German as their mother tongue. A marginally significant difference was also found regarding the listening skills estimated by the parents in the LHQ-3 (*p* = .06). Parents indicated their children's skills on a scale from 1 (very poor) to 7 (excellent). Parents from children of group L1 indicated higher listening skills (*M* = 6.83, *SD* = 0.39) than parents from children of group L2 (*M* = 6.08, *SD* = 1.17). Estimation of the German speaking skills by the parents did not differ significantly between the two groups (*p* = *.183)*.

Parents were asked to indicate the social and emotion regulation skills of their children in the SRS, AQCP and ERSSQ questionnaires. None of the children showed noticeable problems regarding those skills. Furthermore, a Mann Whitney test for the SRS (*p* = .083) and AQCP (*p* = .44) revealed no significant difference between the two groups. The conducted t-test for the ERSSQ did not indicate a significant difference between the two groups (*p* = .659). The results revealed that the groups did not differ in terms of social communication abilities, alexithymia, nor emotion regulation and social skills.

The heart rate and answers to the SAM scale were used as indicators whether frustration has been provoked within the children or not. This is relevant since we are testing the hypothesis that an emotion-eliciting situation has a higher negative impact on the cognitive performance of children when the language used is their first language than when the language used is their second language.

A paired samples t-test for the heart rate of the children during the 5 minutes of resting time and after the frustration eliciting situation showed no significant difference between these two points in time (*p* = .502). However, only for 18 of the 24 children the heart rate measurement worked correctly. These results

indicate that there has not been any change of emotion within the children after receiving their least preferred item. A Wilcoxon signed ranks test was conducted to see whether the answers to the SAM scale come to the same conclusion. The children indicated significantly lower happiness levels after receiving their least preferred item (p = .021) (MR = 7.33) than after the 5 minutes of resting. Further analysis showed that this only applies to the group of children with German as their second language. Regarding the final score in the memory task, a Mann Whitney test revealed that there is no significant difference between the two groups (U = 45.00, z = -1.62 p = .106) All children reached a score between 10 and 14 (L1: M = 11.83, SD = 1.03 and L2: M = 11.17, SD = 1.53), one third of them reached a score of 12 points. The maximum possible score one could reach was 14 and four items were answered correctly by all participants. We did further analysis to see if there would be any differences if those items were removed, but all the results were identical to the ones mentioned above, only the minimum and maximum of the scores changed to 6 and 10.

To conclude our analysis and to finally answer the research question, a regression analysis was conducted to see whether emotionality of the child, the WNV total score and the group affiliation have an influence on the cognitive performance of the participants. Since results showed that there is a significant difference between the two groups regarding their WNV total score, we included the WNV total score in the regression analysis. Emotionality and group affiliation, German as a first or German as a second language, are part of our hypotheses and therefore need to be included in the regression analysis. However, our hypothesis states that children show worse cognitive performance after an emotion-eliciting situation if the language used is their mother tongue. This bases on the assumption that frustration has been elicited in the children. Results showed that only children in group L2 indicated significantly lower happiness levels. Therefore, not the language, German as first or second language, but emotionality itself could cause a worse cognitive performance. Nonetheless, group affiliation is still the most important comparison factor in this study and was therefore included in the analysis. Even though there was a significant difference in the parent's education levels of the two groups, we did not include it in our regression model, because it is a variable that describes more the parents of the children than the children themself. Results showed that group affiliation did not have an impact on the cognitive performance of our participants (*Table 2*). This indicates that there was no difference in the performance between the two groups. Results, therefore, do not support our hypothesis that children show worse cognitive performance if the language used is their first language. According to our results, emotionality also does not have a significant influence on cognitive performance. Only the IQ of the children had a significant impact on their cognitive performance.

Table 2: Linear model of predictors of cognitive performance to check whether the IQ, emotionality during the memory task and German as a mother tongue or second language (L1/L2) have an effect on cognitive performance.

	b	*SE B*	*β*	*p*
Step 1				
Constant	7.75	1.37		.00
IQ	.04	.01	.53	.008
Step 2				
Constant	8.34	7.95		.002
IQ	.03	.02	.47	.057
Change of emotion measured with SAM	.15	.32	.11	.639
Group L1 or L	-2.27	.68	-.11	.692

Note: R^2 = .279 for Step 1, ΔR^2 = .288 for Step 2 (p < .05)
F (3, 20) = 2.699, p = .073

4. Discussion

The findings of the presented study offer an important contribution to the literature

on the relationship between cognition and language proficiency when faced with an emotionally frustrating situation. The aim of this study was to examine whether there is a difference between the use of the mother tongue or a second language during an emotion-eliciting (frustrating) situation in the subsequent cognitive performance (memory) of children aged between 6 and 12 years.

The analyses are based on research previously described in the introduction. For instance, attention is a limited resource and by consuming cognitive resources, in order to regulate a negative emotional state, cognitive performance is reduced as a result. Namely, focusing attention on emotions results in restricted availability of resources for concentrating on another task (Ellis & Ashbrook, 1988; Meinhardt & Pekrun, 2003, as cited in Wentzel & Miele, 2016). Moreover, Boroditsky et al. (2003) suggests that there is a close relationship between language and cognition and that these constructs influence each other. For instance, Martini et al. (2021) show that in Luxembourg children speaking Luxembourgish at home and for whom the main instruction language is German show a better performance in maths and German reading comprehension. In addition, multilingual people show less emotional reactivity by distancing themselves when someone speaks their non-dominant language (Lindquist et al., 2013). These findings suggest that language and emotions could have an impact on cognitive performance.

Moreover, this evidence led us to formulate the hypothesis that children show worse cognitive performance after an emotion-eliciting situation if the language used is their mother tongue, rather than a second language. If this was the case, then the L1 group should have significantly lower scores in the memory performance task and a higher indication of emotionality compared to the L2 group.

First, we assessed the participants' German proficiency level. The results showed marginally significant differences between the two groups. Namely, the L1 group had a marginally higher score than the L2 group indicating a slightly better German language proficiency for the L1 group with German as their mother tongue. Without having different proficiency levels in the two groups, we could not have analysed if speaking a foreign language has indeed an impact because the language level of the second language would have been equivalent to the mother tongue. As indicated in our sample description, the first language of most children in the L2 group was Luxembourgish, a language that is a mix of German and French (The government of the Grand Duchy of Luxembourg, 2022). Therefore it might be easier for children with Luxembourgish as their first language to also have a high proficiency in German, and thus have a better performance in the SET5-10, than children with a more distinct mother tongue such as Portuguese. Furthermore, children in Luxembourg learn German at an early point of time in school which is why they can achieve high proficiency at an early stage. It is also important to take into consideration that the Language level test (SET 5-10) could have been too easy for some children, as the test was conceptualised for native-speaking children aged between 5 and 10 years and our participants' age range went up to 12 years. Therefore, the same table to score the raw values was used for all children 10 years or older. However, the scores reached did not indicate that the test was in general too easy. Scores ranged from 26 to 77 and the highest possible score would be 80 points. None of the children answered correctly to all items.

Another consideration that should not be forgotten in regard to the language skills of the participants is the answers to the language history questionnaire from the parents. There was no significant difference in the judgement of the parents of both groups regarding their children's German speaking skills. Perhaps the parents of the children from the L2 group do not speak German themselves and therefore were not able to judge the level of their children accurately. Hence, even though the participants of the L2 group

had German as their second language, it could have been that their proficiency level was nevertheless very high and very similar to the language level of children being part of the L1 group.
Next, to check for the emotion regulation abilities of the children, the scores of the parent-report questionnaires were compared. The results demonstrated that the groups did not differ significantly in this area with their scores falling into the normal range of values for these questionnaires. Therefore, one could say that the participants had appropriate emotion regulation skills and that Alexithymia, the inability of recognizing emotions, can be excluded as interpretation for the gained results.
Regarding age, gender, and income of the parents there were also no significant differences between the two groups. However, there was a significant difference between the education levels of the parents of the two groups what could also have an influence on the IQ of the children and therefore the cognitive performance results. We will consider this further along in our discussion.
Even though the children's baseline of their heart rate during the five minutes of rest and their heart rate right after they received their least preferred item were not significantly different, the answers to the SAM scale showed a different result, however this was only the case for the L2 group.
Participants of the L2 group were less happy after receiving their least preferred item than at baseline, so this shows that the manipulation led to more emotional reactivity of the L2 group in comparison to the L1 group.
Some of the children of the L1 group remained calm and certain individuals even seemed amused by the fact that we did not give them their preferred item, but their least preferred one instead. For example, one child assumed that at the end, he/she would get their chosen item and therefore did not show the anticipated emotional response.
On the one hand, the manipulation probably did not work for all participants, because of the children's experience with positive reinforcement in everyday situations. However, one has to point out that the manipulation probably affected the participants of the L2 group to a greater extent, because they experienced more difficulties during the SET5-10 and WNV test in general. The struggles encountered by the L2 group may have made them uncertain about their abilities and this led to negative emotions that later influenced the reaction to the emotion-eliciting situation. One could say that in the L2 group the strong negative emotions dominated rather than previous learned positive reinforcement. On the other hand, positive reinforcement could have an impact in the L1 group. "Positive reinforcement is defined as the provision of a stimulus immediately after a behaviour that results in an increase in the use of that behaviour" (Cooper et al., 2007, as cited in Hardy et al., 2020). The application of positive reinforcement in school settings by teachers and at home by parents usually has the goal to enhance desired child behaviours (Wood et al., 2011). Parents or even teachers at school sometimes use the method of operant conditioning and tell the children for example that they will be allowed to play if they do their homework or that they will receive sweets if they finish cleaning their room. Children quickly learn that if they show the wanted behaviour or accomplish the task imposed by their parents, they will be rewarded either with material things or emotional affection by their caregivers. The puzzle was an easy task and, in all likelihood, children also understood that they accomplished this task even though they did not get their most preferred item immediately after solving the puzzle. Therefore, they expected to get their most preferred item, at least at the end of the study as compensation for their effort and participation, as they have been usually rewarded in past situations when achieving other assignments.

The higher levels of negative emotional states shown by participants of the L2 group could also be explained by the fact that they had a lower SES, also demonstrated by the lower education levels of their parents, in comparison to

participants of the L1 group. For this reason, the children of the L2 group have been maybe less spoiled and have not been receiving as many toys or sweets as the L1 group. This could be a reason why they associated a higher meaning to the presented gifts during the testing than children of the L1 group. In relation to their socio-economic status, children of the L1 group for instance probably had a different type of education or knowledge about the presented items and said for instance formulations like “sweets are not healthy to eat” and therefore did not choose them as their first item and were not excited to get them.
Another aspect to consider is that children of the L1 group did not react with strong negative emotions, because maybe the parents “mentally” prepared them, because by having a higher education level they probably experienced similar testing situations and told them to remain calm during the test and to accomplish the tasks without pressure. In order to alleviate their concerns regarding the experiment they may have already told their children beforehand that negative emotions would be temporarily induced during the test. We always told the parents to not reveal any important information regarding the test to their children and tried to answer their questions not in front of the participants to reduce possible bias. However, many parents were concerned about the task leading to frustration and so the assumption that some parents warned their children about the emotional situation beforehand should be considered.
Adding to that, one needs to consider that the different possible answers for the income of the family just separated the socio-economic status into different groups up to a salary of 8499€. Most families indicated an income of more than 8500€ per month, which made us wonder whether the scale is appropriate for Luxembourg. Maybe we would have seen a bigger difference of the socio-economic status between the two groups with more possible answers in a higher range of income. However, it is likely that families participating in scientific studies and therefore contributing to research are of a higher socio-economic status, since the parents then have a higher connection to research because of their own experiences with it.

By analysing the heart rate data, we discovered that some errors occurred with the measurement of the heart rate for 6 of the 24 participants. The data was either missing or it was only assessed for a short period of time. For instance, the apple watch stopped to track the heart rate after five or six minutes, even though we always made sure that the apple watch was functioning during the testing. As there are crucial values missing, important information is lost and this limits the conclusions and interpretations that can be drawn from the data. Given that our dataset is already relatively small because of the sample size, every single data point counts and thus this lost data can lead to imbalanced and biased observations.

As we previously described in our introduction, reacting to a negative emotion such as frustration with disruptive, interfering responses reduces the quality of performance on activities following frustration (Child & Waterhouse, 1953).
Regarding cognitive performance, our predictions were that emotionality, namely frustration and the respective groups, so whether they spoke German as their first or second language would have an impact on the score of the memory task. A linear regression analysis used to see which of the variables actually had an impact on the score showed that only the IQ score from the WNV test was marginally significant (p = .057). Hence, depending on how high or how low the IQ score of the child was, they performed respectively better or worse in the memory task, indicating the cognitive performance in this case. As expected, participants with a higher IQ also showed a better performance in the memory task, while children with lower IQ displayed the opposite pattern. In particular, participants in the L1 group had significantly higher IQ scores. This difference is relevant as the IQ is an

independent variable, which therefore could have had an impact on the results even if we could have controlled it. Comparisons of the education levels of the two parent groups showed that parents of children of the L1 group had significantly higher levels of education which could explain why the participants had higher IQ scores in that group (cf. statistical analysis, Table 1). The study of Kendler et al. (2015) for instance, found that the environment in which children are growing up affects IQ evaluated in late adolescence and that a part of the children's IQ could be explained by the educational level of their parents. These findings are in line with Davis-Kean's (2005) suggestion that parents' education has an impact on their children's achievement through its influence on the parents' educational expectations and specific parenting behaviours. The findings show that the language used (either first or second language), in our study defined by the group, and frustration, more precisely whether the participants were less happy just before the task than before the frustration paradigm, had no significant impact on cognitive performance, giving us an answer to our research question and demonstrating that IQ was the best predictor for the memory task score in this case. Similar evidence demonstrated that cognitive intelligence (IQ) predicted significantly the Iowa Gambling Task (IGT), which also highly relies on cognitive processes like memory, while emotional intelligence (EIQ) did not influence the performance on this task (Demaree et al; 2009). This example also reinforces the fact that the IQ has a higher influence on cognition than emotions.

However, when assessing these results, one should keep in mind that our questionnaire has some limitations. In the questionnaire regarding the video the scores reached by the participants only differed between 10 and 14 points, meaning that each child answered correctly to at least 10 questions. These results bring up the question whether the questionnaire might have been too easy and therefore was not an appropriate means to measure cognitive performance of children this age. We considered not including the four items that were answered correctly by every child in further analysis. However, comparisons of the total scores and the scores without the 4 items showed that the difference between the two groups stays the same.

4.1 Limitations and Outlook

Although the present study helps to contribute to research on the relationship between cognition, language proficiency and emotions, it is not without limitations. At first one should begin by saying that our sample size, consisting of 24 participants, is rather small and thus the generalizability of the results is limited. For time constraint reasons, we were not able to gather a bigger sample. In addition to this crucial point the analysis showed that a lot of results were only marginally significant, which could have been different with a bigger sample.

Furthermore, the heart rate of 25% of the participants was not or only partially recorded due to some technical issues. Therefore, some crucial data was missing. The significant difference in IQ between the two groups and the only marginally significant difference regarding language proficiency had an influence on the results as well as the level of difficulty of the cognitive task. For further research a cognitive task that has already been proven to be appropriate for this age group should be used.

Furthermore, only the participants of the L2 group reacted to the emotionally eliciting situation, probably because they already experienced more difficulties during the SET5-10 and WNV test. This could have led to negative emotions even before the disappointment paradigm and thus could have influenced the actual measured emotion-eliciting situation later on.

As the manipulation did not work for the L1 group this limited our conclusions of our found results.

An explanation as to why some of the participants were not frustrated as expected when confronted with the disappointment paradigm might be because of the childrens' experience with positive

reinforcement in everyday situations as explained previously.
Another reason for this could be that the parents "mentally" prepared their children before the experiment and told them to remain calm when they would be confronted with an emotion-eliciting situation during the test.
Furthermore, a lot of the participants had a high socioeconomic status and thus might not have been interested as much in sweets or toys as they have already had access to these things in their daily lives. Additionally, the possible answers for the families' income were limited as the maximum salary that the parents could indicate was only 8499€. Most parents participating had an income of more than 8500€ per month. Thus, the scale was probably not appropriate for Luxembourgish households and it would have been better to have a higher maximum income as an option.

Overall, the obtained findings suggest that there is no relationship between language and cognitive performance when confronted with an emotional situation. Our expected result that participants of the L2 group would show a better performance in the memory task, because they can distance themselves from an emotional situation when using their non-native language did not occur. This suggests that multilingual students may not have an advantage regarding cognitive performance in an emotional situation when using their second language. Future research should concentrate on the connection of language, emotions and cognition because there is still much space for exploration on this topic. In addition, previously described limitations should be taken into account. For instance, as the difference in language proficiency between the two groups was only marginally significant, this could have influenced the results and should be noted for future research. If possible, the analyses should be conducted on a bigger sample and include more participants having a different mother tongue than Luxembourgish, as this language's syntax resembles the German syntax a lot and therefore could be too close to the L1 group, influencing the findings as a result. It would probably be even better if future research would only include participants in the L2 group having a mother tongue that differs a lot from German and exclude Luxembourgish as it is too similar to German in order to control for this confounding variable.

Funding: The present study was supported by the Luxembourg National Research Fund (FNR) (13651499).

References

Ayçiçegi-Dinn, A., & Caldwell-Harris, C. L. (2004). Bilinguals' recall and recognition of emotion words. *Cognition and Emotion, 18,* 977-987. https://doi.org/10.1080/02699930341000301

Blair, K. S., Smith, B. W., Mitchell, D. G., Morton, J., Vythilingam, M., Pessoa, L., Fridberg, D., Zametkin, E.E., Drevets, W.C., Pine, D.S., Martin, A., & Blair, R. J. R. (2007). Modulation of emotion by cognition and cognition by emotion. *Neuroimage, 35*(1), 430-440. https://doi.org/10.1016/j.neuroimage.2006.11.048

Bond, M. H. and Lai, T. M. (1986). Embarrassment and code-switching into a second language. *Journal of Social Psychology,* 126: 179–186. https://doi.org/10.1016/0147-1767(92)90016-N.

Boroditsky, L., Schmidt, L. A., & Phillips, W. (2003). Sex, syntax, and semantics. *Language in mind: Advances in the study of language and thought,* 61-79. https://web.stanford.edu/class/linguist156/Boroditsky_ea_2003.pdf

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry, 25*(1), 49-59. https://doi.org/10.1016/0005-791694
90063-9

Brown, R. H., Murray, A., Stewart, M. E., & Auyeung, B. (2021). Psychometric validation of a parent-reported measure of childhood alexithymia: The Alexithymia Questionnaire for Children–Parent (AQC-P). *European Journal of Psychological Assessment*. https://doi.org/10.1027/1015-5759/a000640

Butterworth, T. W., Hodge, M., Sofronoff, K., Beaumont, R., Gray, K. M., Roberts, J., Horstead, S. K., Clarke, S. C., Howlin, P., Taffe, J. R., & Einfeld, S. L. (2014). Validation of the emotion regulation and social skills questionnaire for young people with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 44*(7), 1535-1545. https://doi.org/10.1007/s10803-013-2014-5

Caldwell-Harris, C. L. (2014). Emotionality differences between a native and foreign language: Theoretical implications. *Frontiers in psychology, 5,* 1055. https://doi.org/10.3389/fpsyg.2014.01055

Child, I.L. & Waterhouse, I.K. (1953). Frustration and the quality of performance. III. An experimental study. *Journal of Personality, 21*(3), 298-311.https://doi.org/10.1111/j.1467-6494.1953.tb01773.x

Constantino, J. N., & Gruber, C. P. (2012). Social responsiveness scale second edition (SRS-2): *Manual. Western psychological services (WPS).* https://doi-org.proxy.bnl.lu/10.1177%2F1534508410380134

Costa, A. P., Steffgen, G., & Samson, A. C. (2017). Expressive incoherence and alexithymia in autism spectrum disorder. *Journal of Autism and Developmental Disorders, 47*(6), 1659-1672. https://doi.org/10.1007/s10803-017-3073-9

Davis-Kean, P. E. (2005). The Influence of Parent Education and Family Income on Child Achievement. *Journal of Family Psychology, 19*(2), 294-304. https://doi.org/10.1037/0893-3200.19.2.294

Davis, E. L., & Levine, L. J. (2013). Emotion regulation strategies that promote learning: Reappraisal enhances children's memory for educational information. *Child development, 84*(1), 361-374. https://doi.org/10.1111/j.1467-8624.2012.01836.x

Demaree, H. A., Burns, K. J., & DeDonno, M. A. (2010). Intelligence, but not emotional intelligence, predicts Iowa Gambling Task performance. *Intelligence*, *38*(2), 249-254. https://doi.org/10.1016/j.intell.2009.12.004

Eilola, T. M., Havelka, J., & Sharma, D. (2007). Emotional activation in the first and second language. *Cognition and Emotion*, *21*(5), 1064-1076. https://doi.org/10.1080/02699930601054109

Field, A. (2018). *Discovering Statistics Using IBM SPSS Statistics* (5th Revised edition). SAGE Publications Ltd.

Fox, S., & Spector, P. E. (1999). A model of work frustration–aggression. *Journal of organizational behavior, 20*(6), 915-931. https://doi.org/10.1002/(SICI)1099-1379(199911)20:6<915::AID-JOB918>3.0.CO;2-6

Gergoudis, K., Weinberg, A., Templin, J., Farmer, C., Durkin, A., Weissman, J., Siper, P., Foss-Feig, J., Del Pilar Trelles, M., Bernstein, J. A., Buxbaum, J. D., Berry-Kravis, E., Powell, C. M., Sahin, M., Soorya, L., Thurm, A., Kolevzon, A., & Developmental Synaptopathies Consortium (2020). Psychometric Study of the Social Responsiveness Scale in Phelan-McDermid Syndrome. *Autism research: official journal of the International Society for Autism Research, 13*(8), 1383–1396. https://doi.org/10.1093/hmg/ddab280

Graziotin, D., Wang, X., & Abrahamsson, P. (2015, September). Understanding the affect of developers: theoretical background and guidelines for psychoempirical software engineering. In *Proceedings of the 7th International Workshop on Social Software Engineering* (pp. 25-32). https://doi.org/10.1145/2804381.2804386

Grenkowski, G. (2012). *Chomsky's Modularity Hypothesis–Is There an Innate Language Module?*. Grin Verlag. https://www.grin.com/document/190072

Grosjean, F. (2021). *Life as a bilingual: Knowing and using two or more languages*. Cambridge University Press. https://doi.org/10.1017/9781108975490

Hardy, J. K., & McLeod, R. H. (2020). Using Positive Reinforcement With Young Children. *Beyond Behavior*, *29*(2), 95-107. https://doi.org/10.1177/1074295620915724

Hoffmann, D., Hornung, C., Gamo, S., Esch, P., Keller, U., & Fischbach, A. (2018). *Schulische Kompetenzen von Erstklässlern und ihre Entwicklung nach zwei Jahren*. Luxembourg Centre for Educational Testing, Universität Luxemburg; Service de la Coordination de la Recherche et de l'Innovation pédagogiques et technologiques. https://www.semanticscholar.org/paper/Schulische-Kompetenzen-von-Erstkl%C3%A4sslern-und-ihre-Hoffmann-Hornung/66830e0beeefd055a96c8a7e6b454e9fb03946ec

Kendler, K. S., Turkheimer, E., Ohlsson, H., Sundquist, J., & Sundquist, K. (2015). Family environment and the malleability of cognitive ability: A Swedish national home-reared and adopted-away cosibling control study. *Proceedings of the National Academy of Sciences*, *112*(15), 4612-4617. https://doi.org/10.1073/pnas.1417106112

Lindquist, K. A., Satpute, A. B., & Gendron, M. (2015). Does language do more than communicate emotion?. *Current directions in psychological science*, *24*(2),99-108. https://doi.org/10.1177/0963721414553440

Li, P., Zhang, F., Yu, A., & Zhao, X. (2020). Language History Questionnaire (LHQ3): An enhanced tool for assessing multilingual experience. *Bilingualism: Language and Cognition*, *23*(5), 938-944. https://doi.org/10.1017/S1366728918001153

Martini, S. F., Schiltz, C., Fischbach, A., & Ugen, S. (2021). Identifying math and reading difficulties of multilingual children: Effects of different cut-offs and reference groups. *Diversity Dimensions in Mathematics and Language Learning. Perspectives on culture, education, and multilingualism*, 200-228. https://library.oapen.org/bitstream/handle/20.500.12657/50212/9783110661941.pdf?sequence=1#page=217

Miele, D. (2009). *Handbook of motivation at school* (Vol. 704). K. R. Wentzel, & A. Wigfield (Eds.). New York, NY: Routledge. https://doi.org/10.4324/9780203879498

Morton, J. B., & Harper, S. N. (2007). What did Simon say? Revisiting the bilingual advantage. *Developmental science*, *10*(6), 719-726. https://doi.org/10.1111/j.1467-7687.2007.00623.x

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, *25*(1), 49-59. https://doi.org/10.1016/0005-7916(94)90063-9

Petermann, F. (2012). *Sprachstandserhebungstest für Kinder im Alter zwischen 5 und 10 Jahren: SET 5-10* (Vol. 3). Göttingen: Hogrefe.

Quinteros Baumgart, C., & Billick, S. B. (2018). Positive cognitive effects of bilingualism and multilingualism on cerebral function: A review. *Psychiatric Quarterly*, *89*(2), 273-283.https://doi.org/10.1007/s11126-017-9532-9

Richards, J. M., & Gross, J. J. (2000). Emotion regulation and memory: the cognitive costs of keeping one's cool. *Journal of personality and social psychology*, *79*(3), 410. https://doi.org/10.1037/0022-3514.79.3.410

Saarni, C. (1984). An observational study of children's attempts to monitor their expressive behavior. *Child development*, 1504-1513. https://doi.org/10.2307/1130020

Set 5-10 - sprachstandserhebungstest für kinder im Alter Zwischen 5 und 10 jahren – hogrefe verlag. Hogrefe. (n.d.). Retrieved June 6, 2022, from https://www.testzentrale.de/shop/sprachstandserhebungstest-fuer-kinder-im-alter-zwischen-5-und-10-jahren.html

The government of the Grand Duchy of Luxembourg (2022). *An intro to Lëtzebuergesch.* https://luxembourg.public.lu/en/society-and-culture/languages/introduction-letzebuergesch.html

Ugen et al. (2021). Lernstörungen Im Multilingualen Kontext: Diagnose Und Hilfestellungen. Esch-sur-Alzette: *Melusina, 2021.* https://www.melusinapress.lu/read/introduction-learning-disorders-in-a-multilingual-context-a-challenge/section/9c9d3dea-7491-439c-a2ac-cfa2a37d55ae#index.xml-body.1_div.1

WNV - wechsler nonverbal scale of ability – hogrefe verlag. Hogrefe. (n.d.). Retrieved June 6, 2022, from https://www.testzentrale.de/shop/wechsler-nonverbal-scale-of-ability.html

Wood, B. K., Ferro, J. B., Umbreit, J., & Liaupsin, C. J. (2011). Addressing the challenging behavior of young children through systematic function-based intervention. Topics in Early Childhood Special Education, 30(4), 221-232. https://doi.org/10.1177/0271121410378759

Impressum

Luxemburger Experimentalpraktikum Journal
Band 16, Jahrgang 2022

Herausgeber

Dr. Andreia Costa
Koordinatorin des Practical Training in Empiricial Research
Université du Luxembourg
Maison Sciences Humaines
11, Porte des Sciences
L-4366 ESCH

Redaktion

Clémentine Offner, Studentin des BAP

Mitherausgeber

Dr. Caroline Hornung
Dr. Mila Marinova
Talia Retter
Dr. Angelika Dierolf
Dr. Marian van der Meulen
Dr. Ineke Pit-Ten Cate
Dr. Mireille Krischler
Dr. Philipp Sischka
Christina Reisinger
Miriam Zimmer
Dr. Annika Lutz
Dr. Andreia Costa
Maïte Franco
Louise Charpiot
Sam Bernard
Lynn Erpelding

Manuskriptrichtlinien

Die Beiträge richten sich nach dem „Publication Manual of the American Psychological Association" (2010) bzw. den „Zur Gestaltung von Haus- und Abschlussarbeiten" (2004) der Deutschen Gesellschaft für Psychologie.

Erscheinungsweise

Ein- bis zweimal jährlich

Bezug

Ass.-Prof. André Melzer, Université du Luxembourg, Maison Sciences Humaines 11, Porte des Sciences, L-4366 ESCH

Online unter

https://bap.uni.lu

ISSN 3093-1045