Measuring the Approximate Number System

Jomard Sabri

Supervisors
Anders Winman & Marcus Lindskog

Reviewer
Per Andréasson
Abstract

Recent theories in numerical cognition suggest that humans are equipped with a mental system that supports the representation and processing of symbolic and nonsymbolic magnitudes, called the Approximate Number System (ANS). Prior research also suggests that the acuity of the ANS can predict individuals’ mathematical ability. However, results from research within the field has proven to be inconsistent with one another which raises questions about the reliability and validity of methods used to measure the ANS. The present study attempts to replicate the results found in studies suggesting that ANS acuity correlates with mathematical ability. The study also investigates the reliability and validity of different task that have been used to measure the ANS, and also presents a new method of measuring the ANS with an adaptive method. The results show that two tasks correlate significantly with mathematical ability, and multiple regression analyses show that ANS acuity can predict mathematical ability when controlling for general intelligence. Furthermore, the results also further highlight the issue of methodological flaws in previous studies.

Key words: Approximate number system, mathematical ability, methodological inconsistencies
Table of Contents

1 Introduction........................................................................................................................................4
  1.1 The Approximate Number System .................................................................................................4
  1.2 Methodological inconsistencies? .......................................................................................................5

2 Method ................................................................................................................................................11
  2.1 Participants .....................................................................................................................................11
  2.2 Tasks .............................................................................................................................................11
    2.2.1 Numerosity comparison ..............................................................................................................11
    2.2.2 Numerical distance effect ..........................................................................................................11
    2.2.3 Inspection time ...........................................................................................................................12
    2.2.4 Mathematical task ......................................................................................................................12
    2.2.5 Raven’s matrices .........................................................................................................................13
  2.3 Procedure .....................................................................................................................................13
  2.4 Design ..........................................................................................................................................14

3 Results ...............................................................................................................................................14
  3.1 Reliabilities of tasks ........................................................................................................................14
  3.2 Correlations between tasks ............................................................................................................15
  3.3 Multiple regression analyses ...........................................................................................................17

4 Discussion .........................................................................................................................................17
  4.1 The reliabilities of the ZEST-based tasks .........................................................................................17
  4.2 The relationship between direct and indirect measures ...................................................................18
  4.3 The relationship between ANS acuity and mathematical ability ....................................................19
  4.4 The relationship between inspection time and the other tasks .......................................................20
  4.5 Additional findings ..........................................................................................................................20
  4.6 Limitations of the current study .......................................................................................................21
  4.7 Future studies ................................................................................................................................22

5 References .........................................................................................................................................24
1 Introduction

1.1 The Approximate Number System

The ability to deal with numerosities is a fundamental perceptual skill that is used in everyday life. We are constantly exposed to different quantities of objects during our day and we rely on approximating the number of objects when it is not possible to serially count the amount. Researchers have put forth the notion that we use a specific mental system to process and form representations of the number of objects, called the Approximate Number System (ANS). The ANS has been found in human infants (Xu, Spelke, Goddard, 2000), children (Halberda, Feigenson, 2008), adults across cultures (Pica, Lemer, Izard, Dehaene, 2004) and also in nonhuman animals (Feigenson, Dehaene, Spelke, 2004). Some researchers have concluded that the system developed in the earlier stages of the evolution of species that developed it (Barth, La Mont, Lipton, Dehaene, Kanwisher & Spelke, 2006). From an evolutionary point of view, the ANS was developed and used to aid us when determining which tree contains the most fruit, or what group of fish is larger than the other. The physiological basis of the system is the intraparietal sulcus located on the lateral surface of the parietal lobe. fMRI studies have shown that when children and adults are exposed to a nonsymbolic numerosity in different forms (objects, sounds) individuals activate neurons in the intraparietal sulcus (Castelli, Glaser, Butterworth, 2006; Piazza, Price, Butterworth, 2006). Another fMRI study has revealed results showing that children suffering from developmental dyscalculia have lower activity in the intraparietal sulcus when approximating numerosities by comparing magnitudes of objects. In particular, the left intraparietal sulcus and other areas seem to play a crucial role since activity was shown to be correlated with accuracy rates in those areas (Kucian, Loenneker, Dietrich, Dosch, Martin, von Aster, 2006).

Further research of the ANS in human infants and adults has produced interesting results. Research within the domain has suggested that its acuity is related to mathematical achievement (Halberda, Mazzocco, Feigenson, 2008; Mundy & Gilmore, 2009; De Smedt, Verschaffel, Ghesquière, 2008) and that the ANS cannot be trained (Lindskog, Winman & Juslin, 2012). Recent research has also suggested that children with mathematical learning disabilities, such as developmental dyscalculia, also have an impaired ANS, which further bridges the link between the ANS and the intraparietal sulcus. (Mazzocco, Feigenson, Halberda, 2011).
The extensively cited article by Halberda, Mazzocco and Feigenson (2008) in Nature presents interesting results. In their study a total of 64 14-year olds, whose performance in mathematical and general cognitive tasks had been measured longitudinally, were examined and their scores from a variety of tasks were analyzed. This was done to examine individual differences in ANS acuity, and whether these differences correlate with individual differences in cognitive capacities. They found that these individual differences correlate with mathematical ability. They also tested and searched for correlations between a variety of other tasks and the ANS acuity, but the only found significant correlation was between ANS acuity measured at age 14 and the symbolic maths test from Test of Early Mathematics Ability 2nd ed. (TEMA-2) and the Woodcock-Johnson Tests of Cognitive Abilities calculation subtest (WJ-Rcalc) from kindergarten to sixth grade.

A recent study (Inglis, Attridge, Batchelor, & Gilmore, 2011) has suggested that the relationship between the ANS and mathematical ability found by Halberda, et al., (2008) might not be direct, and that the cause of the correlation is due to the ANS and mathematical ability still being in a developmental stage, perhaps by mutually reinforcing one another. This criticism is reasonable since the participants in the previous study were 14 years of age (Halberda & Feigenson, 2008). To test their hypothesis they tested groups of both children and adults. After finding a correlation between ANS acuity and mathematical ability in the child group but not in the adult group it was suggested that the relationship is not as direct as previous researchers had expressed.

1.2 Methodological shortcomings?

Throughout the time researching the ANS, different methods have been used to measure its acuity. One method that has been used to assess the ANS is to measure the numerical distance effect (NDE) in individuals processing representations of numerical magnitudes (used in Peters, et al., 2008). The NDE is an effect that is elicited when individuals compare two different numerocities, either symbolically (e.g. numerals) or nonsymbolically (e.g. arrays of dots). When the ratio between the two amounts is lower (e.g. 9:10) individuals tend to need more time to determine which set of objects has the larger or smaller quantity. When ratios are higher (e.g. 1:2) individuals need less time to determine which quantity is larger or smaller. This
also works with using symbolic stimuli, such as numbers. When asked to determine which number is greater it takes longer time for individuals to make a decision when the numbers are near each other (e.g. 5 and 6) than when the numbers are further from each other (e.g. 1 and 9). In a study measuring the NDE in children with developmental dyscalculia it was shown that children with this impairment have a longer processing and reaction time than children in control groups (Ashkenazi, Mark-Zigdon, Henik, 2009). Together with the previous study by Mazzocco, et al., (2011) showing that children with developmental dyscalculia also have an impaired ANS adds to the suggestion that there is a link between the NDE and the ANS. This link is further strengthened by the found link between the intraparietal sulcus and developmental dyscalculia (Kucian, et al., 2006).

In a study designed to measure the relationship between the tasks previously used in studies to measure the ANS researchers have found that the results from these tasks do not correlate with one another (Gilmore, Attridge & Inglis, 2011). The investigated tasks were nonsymbolic numerosity comparison, nonsymbolic addition, symbolic numerosity comparison and symbolic addition. To test the reliabilities of the tasks’ measurement of the ANS the Spearman-Brown prediction formula was used. The reliability of each task was described as acceptable or good. The researchers ultimately concluded that the participants’ performance and results in these tasks are uncorrelated, despite the results being consistent with theoretical predictions and previous research. They also suggest that these results highlight the methodological flaws in tasks that have been used to assess and measure the ANS. Furthermore, they suggest that it might not only be the ANS that is used when performing these tasks.

Another study conducted by Holloway and Ansari (2009) showed that the NDE elicited from the symbolic version (using Arabic numerals) of the task and the NDE elicited from the nonsymbolic version (using arrays of squares) do not correlate. In another a study designed to challenge the reliability and validity of the NDE researchers have suggested that the mechanisms that are at work when eliciting symbolic NDE and nonsymbolic NDE might not be the same since the results from the two tasks consistently do not correlate. These findings suggest that the nonsymbolic and the symbolic NDE differ in terms of reliability and that they might not measure or tap in to the same system or construct. Furthermore, they suggest that this implies that these tasks measure different cognitive processes when comparing numerical magnitudes. They also propose that when comparing nonsymbolic
magnitudes individuals must create a representation of the magnitude and then approximate it before making a decision. When presented symbolic stimulus, the exact magnitude is already known, and individuals might skip the step of representation and approximation (Maloney, Risko, Preston, Ansari, Fugelsang, 2009).

Furthermore, a study highlighting the methodological inconsistencies when measuring the ANS in infants and children/adults points out that when measuring the ANS in infants, a looking-time paradigm is used. In this paradigm the acuity of the ANS is measured by recording looking time variation when the stimulus changes enough for infants to perceive a difference. When studying children or adults a discrimination paradigm is used. In this paradigm participants are mature enough to direct attention to the stimulus and determine which numerosity is larger. They also highlight the issue that researchers tend to assume that the two different task paradigms measure the same system in similar ways, since results from infant and child or adult studies are often compared. The results from the study show that the much better performance in the discrimination paradigm compared to performance in the looking-time paradigm suggests a considerable difference in task difficulty. In conclusion, the study highlights the fact that either the performance of infants has been underestimated or that of adults overestimated and that it challenges the current ideas about the mechanisms of the ANS. The study also replicated the correlation between ANS acuity and mathematical ability, although addition appears the only significant contributing factor (Gebuis & van der Smagt, 2011).

In other studies, often more recent, the acuity of the ANS is measured by an individual’s internal Weber fraction (\(w\)). The Weber fraction is based on Weber’s law, which states that as the ratio between the numerosities of two stimulus increases, the easier it will be to perceive difference between the two stimuli. To understand its application it is necessary to model the ANS. It is widely accepted that each numerosity is is mentally represented by a distribution of activation on an internal "number line". Since the individual approximates, and does not know the actual magnitude of the representation, this distributed activation is "noisy" and not exact. Every individual has their own acuity in the ANS, and therefore their own activity
patterns on this number line. The Weber fraction is used to quantify the ANS acuity of an individual. Individuals with higher ANS acuity score lower ws.

In the current study, it is argued that tests that elicit effects such as the NDE are a more indirect way of measuring the ANS acuity, which has been suggested in a previous study as well (Price, Palmer, Battista, & Ansari, 2012). This is because tasks that elicit the NDE actually measure one’s reaction time differences, and link these differences to ANS acuity, on theoretical grounds. Tests that calculate one’s w on the other hand measure the discrimination between nonsymbolic magnitudes more directly. The ANS acuity is here quantified by w. The content validity of these tests are seen as higher since they more directly measure ANS acuity. Studies measuring the ANS have used different tasks and measurements without there being much research done on comparisons of these tasks and the correlation between them. In the current study the correlation between the tasks will be investigated. If the tasks, as assumed, do measure the same mechanisms the correlations should be high between the direct and indirect tasks. If these correlations are not found it would be worrying for the research in this field. It would also call in to question the construct validity of the tasks.

The reliability of the method used in Halberda, et al. (2008), is also called in to question in the current study. In the study by Lindskog, et al. (2012) the test-retest reliability of the test used by Halberda, et al. (2008) was investigated. The first reliability test was of two blocks of 100 trials. The correlation was found to be weak and positive but significant with \( r = .35 \) and \( p = .026 \). When testing the reliability with two blocks of 500 trials each the correlation was strong, positive and significant with \( r = .79 \) and \( p < .001 \). The reliability of the test was shown to increase with the amount of trials, as expected from classical test theory. To further increase reliability this study will be using the same method used in Lindskog, et al. (2012), however with an adaptive method. The trials generated in the Lindskog, et al. (2012) study were random and not reliant on whether the participant was correct or incorrect in determining the larger numerosity in the previous trial. This study aims to investigate whether an adaptive method could increase the reliability of the test. The method used in the study by Halberda, et al., (2008) is also an example of a non-adaptive method.

---

1 For more detailed information about the Weber fraction and its application in measuring the ANS, visit: http://www.panamath.org/wiki/index.php?title=What_is_a_Weber_Fraction%3F#Model_Representations_of_the_ANS
The trials in that method were generated without relation to whether the participant was correct or incorrect on the previous trial. The adaptive method would be included so that every generated trial is based on whether the participant was correct or incorrect on the previous trial. When correct, the adaptive method would generate a dot array of equal or slightly higher difficulty (lower ratio of dot arrays). When the participant is incorrect the next trial generated would be easier (higher ratio). This adaptive method is based on the findings presented in the article by Watson and Pelli (1983). In this article the authors argue for the advantages of the adaptive method since it generates response-based trials in real-time. This method is more efficient in measuring an individuals threshold ($w$ in the current study). This procedure of generating the next trial has been dubbed QUEST by its authors. The procedure used in the current study is however based on the ZEST procedure, a modification of the QUEST procedure. It has shown to be a more reliable method (King-Smith, Grigsby, Vingrys, Benes, Supowit, 1994).

The study is also designed to measure participants information processing ability with the inspection time (IT) task. This task measures individuals’ “mental processing speed” and studies have shown that the results from the task accounts for about 20% of intelligence-test variance. It is also suggested that there is a strong association between inspection time and psychometric intelligence (Deary, Stough, 1996; Deary, Simonotto, Meyer, Marshall, Marshall, Goddard, Wardlaw, 2004). The IT task involves being exposed to the stimulus for very short amounts of time. The participant is required to quickly perceive the stimulus and make an either-or decision. The task is perceived as very difficult by the participant. Furthermore, the task is included in this study to investigate the extent of which acuity in one’s ANS is related to the ability to quickly perceive and process stimulus, since both tasks involve being exposed to stimulus for short amounts of time (Lindskog, Winman, Juslin, 2012).

Since the participants in this study performed several tasks we will be able to look for correlations between the results of the different tasks. A task involving symbolic numerosity comparison will also be used to measure the ANS through the NDE. Any found correlations would further our understanding in the measuring of the ANS and the convergent reliability of these methods.
With this design this study aims to investigate the relationship between individuals’ ANS acuities and mathematical ability. Four different tasks are used to measure the ANS: the Halberda, et al., (2008) nonsymbolic numerosity comparison (non-adaptive), the ZEST-based nonsymbolic numerosity comparison (sequential presentation), the ZEST-based nonsymbolic numerosity comparison (simultaneous presentation) and the symbolic numerosity comparison (NDE). The three remaining tasks are the inspection time task, the mathematical task and Raven’s matrices.

The aims of the current study are:

1. Previous research has suggested that higher reliability is difficult to achieve when measuring procedures are not lengthy enough (too few trials). This study aims to investigate the reliability of the measuring procedure with an adaptive method based on ZEST.
2. What is the correlation between direct and indirect measurements of the ANS? If these tests measure the same system, which is assumed, correlations should be found and be high.
3. To determine the validity of the tasks measuring the ANS, the correlation with mathematical ability will also be examined. If a significant correlation is found the predicting power will be examined when controlling for intelligence.
4. Does inspection time correlate with scores from other tasks?
2 Method

2.1 Participants

40 participants (27 female, 13 male) took part in the study. Most participants were students from Uppsala University with a mean age of 24.6 years (SD=8.2 years). Participation in a study is a compulsory assignment for students studying psychology at the institution and those who participated for that reason received a certificate of participation. The other participants received a ticket to a movie theater after participation.

2.2 Tasks

2.2.1 Nonsymbolic numerosity comparison. In the nonsymbolic numerosity comparison task participants were presented arrays of blue and yellow dots. When ready, participants pressed a button on the keyboard to present the stimulus, which was exposed for 300 ms to prevent the participant from serially counting the magnitude. Three different versions of the task were used. The first version was the same method as used in the study by Halberda, et al., (2008) which presents dot arrays simultaneously and is non-adaptive. In the second version of the task, which is based on the adaptive method, the dot arrays were presented sequentially, separated by a 300 ms blank interval. In the third version, also based on the adaptive method, the dot arrays were presented simultaneously. The number of dots varied between 11 and 31. The adaptive test selected the closest ratio between the dots matched on the current estimate of w. Based on previous results (Lindskog, et al., 2012) a w of .23 was used as prior starting value. Participants recorded their answer by pressing one of two buttons on the keyboard. One for when the blue dots were perceived to be of larger magnitude, and one for when the yellow dots were perceived as so. The better the measured acuity, the smaller the Weber fraction would be. The method of calculating participants’ Weber fractions can be found in Appendix 1. The test with simultaneous presentation will be referred to as “ANS-Sim”, the test with sequential presentation as “ANS-Seq” and the test based from the study by Halberda, et al., (2008) as “ANS-Hal”.

2.2.2 Symbolic numerosity comparison (NDE). The numerical distance effect was measured in participants using a program that for each trial presents a fixation point and afterwards presents two single-digit numbers and asks the participant to choose what number is larger as quickly as possible. There was a fixation point for
500 ms and then the two numbers were simultaneously presented to the left and right of the fixation point. The numbers were visually present until the participant had answered which number was larger. The size of the effect was measured by comparing the reaction times for “near” (e.g. 5-6) and “far” (e.g. 1-9) combinations of numbers, aggregating them in to an average reaction time and then the difference was calculated with the two values. The size was calculated by subtracting the average RT for far trials from the average RT for near trials. This task will be referred to as “the NDE task” further on.

2.2.3 Inspection time task (IT). Inspection time is a two-alternative visual backward-masking task. On each of the 100 trials in the task participants first saw a fixation cross for a duration of 500 ms. The fixation cross was followed by a stimulus with one horizontal line and two vertical lines organized to somewhat resemble the Greek letter π. The two horizontal lines were of different lengths and participants had to discriminate which of the two was the longest. The π-stimulus was presented with one of five presentation times (25, 40, 60, 80, 100 ms) and participants saw 20 presentations from each presentation time. The π-stimulus was followed by a 500 ms mask over the two horizontal lines. Participants gave their answer by pressing one of two keyboard buttons, one for when the left line was perceived as longer, and one for when the right line was perceived as longer. Participants’ results were based on the number of correct answers, with the maximum being 100 correct answers.

2.2.4 Mathematical task. To measure participants’ mathematical ability a test based on the mathematical task found in the study by Gebuis & van der Smagt (2011) was used. Participants performed four different sets of mathematical problems: addition, subtraction, multiplication and division. The task started with simple problems with difficulty increasing with the number of problems solved. To increase difficulty more digits were used and the requirement of carrying or borrowing resulted in the gradual increase in difficulty to calculate: addition (e.g. 2+7, 13+8, 21+34), subtraction (e.g. 9-6, 17-13, 43-28), multiplication (2*4, 8*5, 12*6) and division (8/2, 16/4, 54/3). Participants were given a time limit of two minutes and 30 seconds (150 seconds) for each subtest and asked to solve as many problems as accurately as they could within the time limit. There was a short break between the subtests. The results from this test have been split in five parts. One point was awarded for each correct answer. The total amount of correct answers will be referred to as “MTSum” scores.
2.2.5 Raven’s matrices. Raven’s matrices is a non-verbal multiple choice test designed to measure the reasoning component of one’s general cognitive functioning (fluid intelligence). In this test participants were shown a 3 by 3 matrix of patterns and asked to determine and choose the correct missing part of the pattern from a pool of eight choices. Participants had a time limit of 15 minutes to solve as many patterns as possible. Raven’s matrices is one of the most commonly used tests to measure general cognitive ability, and loads high on a fluid intelligence factor. This task is included in the current study to be used in eventual multiple regressions, to control if ANS-measuring tasks still significantly predict mathematical ability when intelligence is controlled for. This version of the test is based on the version used in the study by Stanovich & West (1998). This version is a shortened version of the original task, which has 30 problems that need to be solved. The shortened version does not involve the 12 easiest problems, in which performance in a college sample is near ceiling, and has also excluded the six hardest problems in which performance is nearly floored (Stanovich & West, 1998). For every correct answer, participants received one point, with the maximum amount of points being 18.

2.3 Procedure

All participants performed the different tasks on a computer with a different program for each task. The order of performance was 2x120 trials of the ANS-Sim task, 2x120 trials of the ANS-Seq task, 100 trials of the ANS-Hal task, based on Halberda, et al., (2008), 160 trials of the NDE task, 100 trials of the inspection time task, the mathematical task, another 160 trials of the NDE task, another 100 trials of the inspection time task and finally Raven’s matrices. The reason for splitting most of the tests was to calculate the test-retest reliability of the tests, which will be corrected using the Spearman-Brown prediction formula.

Before each task participants were guided through a short test round in order to familiarize them with the task at hand to avoid the learning phase of each task and get as accurate readings as possible.

The total amount of time used to perform all the tasks varied between and hour and a half and two hours. After completing all the tests participants were asked to judge how motivated they were to perform their best throughout the participation, on a scale from 1 (not motivated) to 10 (very motivated). Information about age and gender was collected before participation in the tasks.
All participants were asked to read a notice informing them of their right to abort participation at any time, that they would remain anonymous in the data collection and that they would not be exposed to arousing or emotionally charged stimuli.

2.4 Design

To find any potential relationships between the different variables mentioned the design of this study is a non-experimental correlation analysis with eventual multiple regression analyses. Any correlation found will be used to theorize and shed light about the relationship between variables. The reliabilities of most tasks will also be calculated and multiple regression analyses will be conducted to see whether the ANS-measuring tasks can still significantly predict MA when controlling for general intelligence, through Raven’s matrices.

3 Results

To generate results for the study all correlations between the variables were calculated, in line with previous research. When scanning for outliers (± 3 standard deviations) none were found and the results generated are based on the scores of 40 participants. The reliability of some tasks were also calculated, with focus on the reliability of the ZEST-based nonsymbolic comparison tasks (ANS-Sim and ANS-Seq).

3.1 Reliabilities of tasks

As previously mentioned some tasks were split up in half so that the test-retest reliability of the tasks could be calculated. Participants’ scores from the first half were compared to the scores of the second half in order to measure the correlation between the scores and the tasks’ reliability. Since the reliability of the tasks were based on only half the amount of the total amount of trials of the aggregated full measures, they were corrected using the Spearman-Brown prediction formula. The reliabilities appear in Table 1.
Table 1. *Test-retest reliabilities and respective corrected reliabilities by Spearman-Brown correction for the aggregated measures.*

<table>
<thead>
<tr>
<th></th>
<th>Reliabilities</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test-retest</td>
<td>Spearman-Brown</td>
<td></td>
</tr>
<tr>
<td>ANS-Sim</td>
<td>.58</td>
<td>.74</td>
<td></td>
</tr>
<tr>
<td>ANS-Seq</td>
<td>.85</td>
<td>.92</td>
<td></td>
</tr>
<tr>
<td>ANS-Hal</td>
<td>.35</td>
<td>Not available</td>
<td></td>
</tr>
<tr>
<td>IT</td>
<td>.84</td>
<td>.91</td>
<td></td>
</tr>
<tr>
<td>NDE</td>
<td>.66</td>
<td>.79</td>
<td></td>
</tr>
</tbody>
</table>

All calculated reliabilities based on df=38.

3.2 Correlations between tasks

Before checking for correlations between tasks the participants’ scores were aggregated in order to have one value for each task. The correlations between these aggregated scores were analyzed. Because reliabilities are known, values corrected for attenuation due to unreliability were calculated and appear within parentheses in Table 2. Several interesting significant correlations were found.

Table 2. *Correlations between all tests used in the study.*

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANS-Sim</td>
<td></td>
<td>.34*</td>
<td>(.41*)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANS-Seq</td>
<td></td>
<td>.48**</td>
<td>(.94**)</td>
<td>.00</td>
<td>(.00)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANS-Hal</td>
<td></td>
<td>-.10</td>
<td>(-.12)</td>
<td>-.14</td>
<td>(-.15)</td>
<td>.17</td>
<td>(.30)</td>
<td></td>
</tr>
<tr>
<td>IT</td>
<td></td>
<td>.43**</td>
<td>(.56**)</td>
<td>.20</td>
<td>(.23)</td>
<td>.03</td>
<td>(.05)</td>
<td>-.37*</td>
</tr>
<tr>
<td>NDE</td>
<td></td>
<td>-.32*</td>
<td>(-.32*)</td>
<td>-.02</td>
<td>(-.02)</td>
<td>-.28</td>
<td>(-.47*)</td>
<td>.10</td>
</tr>
<tr>
<td>MTSum</td>
<td></td>
<td>-.25</td>
<td>(-.31)</td>
<td>-.22</td>
<td>(-.27)</td>
<td>-.09</td>
<td>(-.16)</td>
<td>.19</td>
</tr>
<tr>
<td>Raven's</td>
<td></td>
<td>.47**</td>
<td>(-.54**)</td>
<td>-.37*</td>
<td>(-.38*)</td>
<td>-.20</td>
<td>(-.34*)</td>
<td>.17</td>
</tr>
<tr>
<td>Motivation</td>
<td></td>
<td>.47**</td>
<td>(-.54**)</td>
<td>-.37*</td>
<td>(-.38*)</td>
<td>-.20</td>
<td>(-.34*)</td>
<td>.17</td>
</tr>
</tbody>
</table>

*p<.05, **p<.01. Correlations within parentheses are values after correction for attenuation due to unreliability. All correlations based on df=38.

The mathematical task was also split in to the four parts that it is made up of: the addition part, subtraction part, multiplication part and the division part. This was done to investigate whether tasks correlated significantly with at least parts of the mathematical task.
Table 3. Correlations between tests used to measure ANS acuity and scores on the mathematical task.

<table>
<thead>
<tr>
<th></th>
<th>Correlations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Addition</td>
</tr>
<tr>
<td>ANS-Sim</td>
<td>-.43** (.50**)</td>
</tr>
<tr>
<td>ANS-Seq</td>
<td>-.07 (.07)</td>
</tr>
<tr>
<td>ANS-Hal</td>
<td>-.32* (.55*)</td>
</tr>
<tr>
<td>NDE</td>
<td>-.25 (.28)</td>
</tr>
</tbody>
</table>

* p<.05, ** p<.01

Correlations within parentheses are values after correction for attenuation due to unreliability.

All correlations based on df=38.

When splitting the mathematical task into the subtasks, significant correlations were found between the addition and subtraction part of the mathematical task and the ANS-Hal task. In fact, correlations (corrected for attenuation) between these tasks were the strongest found in the study (addition: \( r(38) = -.32 (.55), p=.04 \), subtraction: \( r(38) = -.39 (.66), p=.01 \)). As expected, the ANS-Sim task correlates significantly with MTSum, as well as with the addition part \( (r(38) = -.43 (.50), p<.01) \) and the subtraction part \( (r(38) = -.31 (.36), p=.05) \). The correlation with the multiplication part is close to significance \( (r(38) = -.30, p=.06) \), which could explain why the ANS-Sim task correlated with the mathematical task as a whole, while the ANS-Hal task did not before the correction.

Somewhat surprisingly the NDE task does not correlate significantly with any part of the mathematical task. The same is to be said about the absence of significant correlations between the ANS-Seq task and all parts of the mathematical task.

3.3 Multiple regression analyses

Of the different mathematical subtests, the division task correlates lower with the ANS measures and the other subtests, and thus appears to deviate from the others. Because of this, a composite mathematical ability measure was created excluding division but including addition, subtraction and multiplication, MTComp.

There are four candidate measures of the approximate number system in the study, ANS-Hal, ANS-Sim, ANS-Seq and NDE. These tests should predict mathematical ability performance if previous theory is correct. Multiple regression analyses were performed with each of the above mentioned measure as independent
variable together with the proxy for fluid intelligence, Raven’s matrices and with the MTComp mathematical ability measure as dependent variable.

For NDE the β-weight (-.10) in the regression failed to reach significance \((p=.49)\). For ANS-Seq the β-weight was close to zero (-.001), \((p=.99)\). However, for ANS-Sim the β-weight (-.31) was significant \((p=.045)\) and explaining 17.9% of the variance \((\text{Adj. } R^2 = .179, F(2,37) = 5.25, p = .01)\). Likewise, the β-weight for ANS-Hal (-.3) was significant \((p=.03)\) and explaining 17.7% of the variance \((\text{Adj. } R^2 = .177, F(2,37) = 5.21, p = .01)\).

Thus, neither NDE nor the sequential ANS task predicted mathematical ability whereas both parallel measures of ANS predicted this capacity when general cognitive functioning (fluid intelligence) was controlled for.

### 4 Discussion

This study aims to contribute to the growing body of research focusing on the ANS and the reliability of the tasks used to measure the ANS. To recap, the aims of the study were 1) to investigate the reliability of ZEST-based tasks when measuring the ANS, 2) to investigate the relationship between direct and indirect measurements of the ANS, 3) to investigate whether ANS acuity is correlated to mathematical ability, and whether ANS acuity could predict mathematical ability when controlling for intelligence, and 4) to investigate whether inspection time correlates with scores from the other tasks used in the current study.

#### 4.1 The reliabilities of the ZEST-based tasks

To answer the first question of the study, the reliability of the ZEST-based tasks were measured. The measured reliability of the ZEST-based tasks were higher than the reliability of the ANS-Hal task, which was also used in the study by Halberda, et al., (2008). These findings indicate that the ZEST-based method of generating stimulus is a more reliable method than one that generates dot arrays at random, with no regard to whether the participant was correct or incorrect in the previous trial. This also makes it possible to measure an individuals ANS acuity without long test procedures, ultimately reducing the work load of the participant and the creators of the study. The average completion time for the ANS-Sim task was 7 to 12 minutes. This also makes it possible for future studies to incorporate more tasks in the same study without making participants endure long test procedures, which could affect the
data negatively due to mental exhaustion or inattentiveness. The ANS-Sim tasks reliability of $r(38)=.74$ is based on 240 trials. In the study by Lindskog, et al., (2012), where the ANS-Hal task was used, the reliability of the ANS-Hal task was calculated to $r(38)=.79$, but this value was based on 500 trials. In essence, the reliability of the ANS-Sim task differs by .05 compared to the reliability of the ANS-Hal task, even though the total amount of trials is less than half. The benefits of the ZEST-based method are a welcomed addition to the field.

4.2 The relationship between direct and indirect measures

The second goal of the study was to investigate and theorize about the relationship between direct and indirect measures of the ANS. A significant correlation was found between scores from the ANS-Sim task and the NDE task. In previous research both have been used to measure the ANS. Although this correlation is positive and significant, it is far from as high as expected if the tasks were to measure the same construct or system, even when we adjust the correlation for unreliability with correction for attenuation. It was previously discussed how the NDE task could be seen as a more indirect method of measuring the ANS, while the ANS tasks could be seen as a more direct method because of the way it is measured by a Weber fraction. Because previous research has linked ANS acuity to mathematical ability, it is interesting to note that participants’ score from the NDE task do not significantly correlate with the mathematical task, and not the sub-tasks either, while the ANS-Sim task does. This finding raises questions about the validity and reliability of indirect symbolic methods previously used in research to measure the ANS. However, previous research has shown that children with developmental dyscalculia show worse ANS acuity when measured through the indirect method of eliciting the NDE in participants (Ashkenazi, Mark-Zigdon, Henik, 2009), but also when measured with direct nonsymbolic numerosity comparison tasks, such as the ones used in the current study. (Mazzocco, Feigenson, Halberda, 2011). This further adds to the more and more established notion that the ways of measuring the ANS are not nearly good enough to be able to distinguish the processes that are involved when using direct and indirect tasks. The results presented in the study invite to further research being conducted to determine and investigate additional differences between the two ways of measuring the ANS.
4.3 The relationship between ANS acuity and mathematical ability

To investigate this relationship multiple correlation analyses were made. Table 2 shows the found significant correlations. The ANS-Sim task is significantly correlated with the mathematical task, whereas the ANS-Hal task was not before correction for attenuation, but is after the correction. The ANS-Seq task does not correlate with mathematical ability at all, and is non-significant which is interesting as the study conducted by Gebuis & van der Smagt (2011) found a significant correlation between their discrimination task scores and mathematical ability. The method used by the authors in the discrimination paradigm was one of sequential stimulus presentation, resembling the ANS-Seq task used in the current study. But, as previously revealed, the ANS-Seq task did not correlate significantly with the mathematical task, which is taken from this study. This further adds to the discussion about the reliability of tasks used in studies measuring the ANS. This also coincides with the results from the study by Gilmore, et al., (2011) where no correlations were found between the six tasks that have been used in previous research to determine ANS acuity. The NDE task also fails to significantly correlate with the mathematical task, which goes against the results of some studies that have found the opposite relation between the two (De Smedt, et al., 2008; Holloway & Ansari, 2009).

It is evident that the ANS-Sim and ANS-Hal tasks, which simultaneously present dot arrays, correlate with the mathematical task, whereas the ANS-Seq and NDE task do not. To further investigate the relationships between the tasks and scores on the mathematical task it was split up into the four parts it consisted of. When checking for correlations with the four different parts of the mathematical task it was found that the ANS-Hal task did correlate significantly with the addition and subtraction part of the task. The ANS-Sim task correlates similarly, however the correlations found were weaker, yet more significant. The correlation between the ANS-Sim task and the multiplication part of the mathematical task is very close to significance ($p=0.06$), and is significant after correction for attenuation, which could explain why the ANS-Sim task significantly correlates with MTSum as a whole, when the ANS-Hal task did not (before correction). To conclude, participants’ scores on the ANS-Sim and ANS-Hal tasks were found to correlate with mathematical ability which means that the results from the study conducted by Halberda, et al., (2008) were replicated and that the ANS seems to play a significant part in human numerical cognition.
To further investigate the relationship between ANS acuity and mathematical ability multiple regression analyses were calculated with the simultaneous ANS tasks (ANS-Sim and ANS-Hal) in two models while controlling for intelligence. The beta weights for each task remained significant which indicates that these tasks measure a domain-specific mathematical ability. These results suggest that ANS acuity is somehow and somewhat linked to mathematical ability. This link further suggests that our ability to process numbers and magnitude is linked to the evolutionary ancient system which deals with numerical approximation.

4.4 The relationship between inspection time and the other tasks

Furthermore, a rather unexpected correlation between the IT task and the NDE task was found in the current study. The correlation is of medium strength \((r(38)=-.37(-.43))\) and one possible explanation could be that both tasks involve time restraint in one form or another. The IT task presents stimulus with different exposure times and the NDE task requires participants to answer as fast as possible. The link could also be because of the design of the two tasks. They are similar when it comes to decision-making for the participant. Both tasks present an either-or choice for the participant after the stimulus is exposed. Also, the IT task does not correlate with any ANS task which suggests that the design of the ANS tasks does not significantly measure participants’ skill in quickly perceiving stimulus. Note also that participants performed in the NDE task and then the IT task twice throughout their participation in the study. No other significant correlations were found between the IT task and other tasks, suggesting that the extent of which acuity in one’s ANS is related to the ability to quickly perceive and process stimulus does not affect the reliability of measuring ANS acuity.

4.5 Additional findings

Additionally, the correlation found between ANS-Sim and ANS-Seq suggests that these tasks somewhat tap in to the same mental system. However, the correlation between ANS-Seq and MTSum is almost non-existent, and is not significant. The ANS-Sim task does however correlate with MTSum. This could mean a number of things. Perhaps these two tasks tap in to the same system, to an extent, but where the simultaneous presentation triggers some form of mathematical processing of the stimulus, whereas the sequential task does not trigger the same activity. Another idea
is that the ANS-Seq task, which presents stimulus sequentially, activates and involves the short-term memory (STM) of the participant (the stimulus were separated by 300 ms and a blank display that prevents iconic memory from being used). This would mean that the ANS-Seq task activates both the ANS and STM, and perhaps even other mental processes, which might not correlate with one another. The ANS-Sim task however does not involve an STM process, since stimulus is presented simultaneously and participants are asked to answer immediately after the presentation. With this task participants are likely to form a representation after the stimulus disappears and when immediately asked to answer, the STM might not be activated or used at all. These reflections could be used to further distinguish the differences between direct and indirect methods of measuring the ANS.

To conclude the discussion directly related to the results of the study, it could be said that these findings further highlight the methodological inconsistencies of previous studies, a concern also voiced in the study by Price, et al., (2012) and Maloney, et al., (2010). An optimal method is yet to be found that measures individuals’ ANS acuity with optimal precision. To make this possible, further research should focus on the psychometrics and methods to further increase the reliability of the tests used. Methods used so far have been proven to be inaccurate by previous studies as well as by the results of the current study. The results from this study also show that the ZEST-based tasks have a higher reliability but there is still work to be done to further improve the reliability, and with it the content validity of the tasks as well. These inconsistencies might be easier to overcome when viewing methods as either direct or indirect. This distinction is still not used and further research should be conducted to further examine the difference between the two paradigms.

4.6 Limitations of the current study

To highlight the shortcomings of the current study several extraneous variables were noted during data collection as they might have affected the data collection process and participants’ efforts. Also, the 40-person sample can not be seen as representative for the population as a majority of participants were university students at the time the study was ongoing.

Some participants also showed less enthusiasm and probably did not perform their best during the testing process. This might be because of the lack of supervision
and that some participants were students who had to participate in a study or two as a compulsory task when studying psychology courses at the institution. The reward, a ticket to the movie theater, might not have been enough of a reward to spur participants for the tests. Even though motivation scores correlated strongly with scores on a number of tests, it can be assumed that when participants were asked how motivated they were their answers were affected by social pressure, as in not wanting to mention how unmotivated they were when in fact they are receiving a reward for participating. Some participants did however show motivation by voicing their appreciation of the challenge and difficulty of some tasks and their scores were found to be above average.

Furthermore, when reading the instructions for the mathematical task, some participants showed obvious signs of nervousness ahead of the task. It seemed that these participants were nervous because they were going to be solving mathematical problems. Several participants voiced their nervousness by stating that it had been a while since they were tested in anything involving mathematical problems. As data collection supervisor, it was often noted that when a participant was nervous ahead of the task their performance would drop compared to other participants.

Also, some participants took as long as two hours to complete all the tasks in the study. This could have affected performance in the latter parts of the tasks as several participants commented on the length of the experiment. Since the IT task and the NDE task were split, with one half of the test in each half of the whole design, the test-retest reliabilities might have been affected by participants becoming increasingly bored or tired with the tasks.

4.7 Future studies

In final consideration, because of the light cast on the methodological inconsistencies by the study, and also the highlighting of results which contradict previous findings, future studies should focus on the issues discussed to enable further and deeper analysis of ANS acuity, its relation to mathematical ability and also the reliability of the methods used to measure ANS acuity. It is apparent that studies so far have been using less reliable and less valid methods to draw bold conclusions. It is however positive to note that the flaws of previous studies are being brought to attention by recent studies (Price, Palmer, Battista, Ansari, 2012). Future studies should also further investigate the relationship between mathematical ability and ANS
acuity, even though this relationship has been extensively researched. However, it is because of the previously mentioned concerns with task reliability that this relationship should still be researched when better methods of measurement are created. Further research in line with this study and the previously cited one is welcomed so that the unknowns of the approximate number system could be revealed and thoroughly researched.
5 References


Barth, H., La Mont, K., Lipton, J., Dehaene, S., Kanwisher, N. & Spelke, E.S. (2006). Non-symbolic arithmetic in adults and young children, *Cognition, 98*, 199-222.


Appendix 1

To model the performance in the ANS acuity task we used a classical psychophysics model that relies on a linear format of the NAS. This has been shown in earlier work (e.g. Halberda, et al., 2008) to be a plausible model of performance in numerical discrimination. Percentage correct was modeled as a function of increasing ratio between the two sets of blue and yellow dots (larger sample/smaller sample, or \( n_2/n_1 \)). In the psychophysics model, each numerosity is represented as a Gaussian random variable with means \( n_2 \) and \( n_1 \), and standard deviations equal to the Weber fraction \( w \). Subtracting the Gaussian for the smaller set from that for the larger set returns a new Gaussian that has a mean of \( n_2 - n_1 \) and a standard deviation of \( w \sqrt{n_1^2 + n_2^2} \). Percentage correct is then equal to \( 1 - \text{error rate} \), where error rate is defined as the area under the tail of the resulting normal curve computed as follows

\[
\frac{1}{2} \operatorname{erfc}\left(\frac{|n_1 - n_2|}{\sqrt{2}w\sqrt{n_1^2 + n_2^2}}\right)
\]

where \( \operatorname{erfc} \) is the complementary error function. This fits percentage correct in the ANS acuity task as a function of the Gaussian approximate number representation for the two sets of dots with \( w \) as a single free parameter. The individual Weber fraction obtained from such a model fit describes the standard deviations for the Gaussian representation of the ANS acuity, thus describing how much the two gaussian representations overlap and thereby predicting an individual percentage correct on a numerical discrimination task. We used this model to find the best fit for each individual separately. A MATLAB implementation used the Levenberg-Marquardt algorithm for a nonlinear least-squares fit on the average percent correct in each ratio bin for each subject. The program tries to find the best fitting \( w \) by attempting to reduce the sum of squared error within 50 iterations.