The Cognitive Underpinnings of Non-symbolic Comparison Task Performance

Sarah Clayton

A thesis submitted for the degree of

Doctor of Philosophy

Loughborough University

February 2016

Abstract

Over the past twenty years, the Approximate Number System (ANS), a cognitive system for representing non-symbolic quantity information, has been the focus of much research attention. Psychologists seeking to understand how individuals learn and perform mathematics have investigated how this system might underlie symbolic mathematical skills. Dot comparison tasks are commonly used as measures of ANS acuity, however very little is known about the cognitive skills that are involved in completing these tasks. The aim of this thesis was to explore the factors that influence performance on dot comparison tasks and discuss the implications of these findings for future research and educational interventions.

The first study investigated how the accuracy and reliability of magnitude judgements is influenced by the visual cue controls used to create dot array stimuli. This study found that participants' performances on dot comparison tasks created with different visual cue controls were unrelated, and that stimuli generation methods have a substantial influence on test-retest reliability. The studies reported in the second part of this thesis (Studies 2, 3, 4 and 5) explored the role of inhibition in dot comparison task performance. The results of these studies provide evidence that individual differences in inhibition may, at least partially, explain individual differences in dot comparison task performance. Finally, a large multi-study re-analysis of dot comparison data investigated whether individuals take account of numerosity information over and above the visual cues of the stimuli when comparing dot arrays. This analysis revealed that dot comparison task performance may not reflect numerosity processing independently from visual cue processing for all participants, particularly children.

This novel evidence may provide some clarification for conflicting results in the literature regarding the relationship between ANS acuity and mathematics achievement. The present findings call into question whether dot comparison tasks should continue to be used as valid measures of ANS acuity.

Contents

List of Figures

List of Tables

Acknowledgements

First and foremost, I would like to thank my supervisors Camilla Gilmore and Matthew Inglis who have provided me with so much valuable guidance and support throughout the duration of my PhD. I am extremely appreciative of the opportunities they have given me and of the huge amount of time and effort they have put in to my development as a researcher. I couldn't have asked for better supervisors.

This research wouldn't have been possible without the help of all the participants who took part in my studies, so a big thank you goes to those who volunteered to compare hundreds of dot arrays for me. In particular I would like to thank Cobden Primary School in Loughborough, and the University of Nottingham Summer Scientist Week team for allowing me to conduct my research with them.

Thank you to Ross Carter and Giles Taylor for their programming expertise, making it possible to quickly join up dot images and measure their convex hulls. I would also like to thank all of the friends I've made in the Midlands Mathematical Cognition Group for the useful discussions, insightful feedback, and of course the wonderful cakes.

Finally, I am incredibly grateful to my friends and family for all their encouragement. A special thanks goes to David for his endless supply of patience and understanding.

Declaration

I, the author, declare that the work presented in this thesis is my own and has not been submitted for a degree at any other institution. The results of Study 1 have been published in *Acta Psychologica* and the results of Study 2 have been published in *ZDM Mathematics Education*. None of the remaining work has previously been published in this form.

Part I

General Introduction

Chapter 1

Literature Review

The aim of this chapter is to present an overview of the literature surrounding non-symbolic comparison tasks in order to provide a background for the empirical work reported later in this thesis. First, I begin with a brief introduction to the underlying cognitive skills that are thought to be important for learning and performing mathematics, including general processing skills and more specific mathematical skills. Next, I narrow my focus to review one particular domain-specific skill in detail, Approximate Number System (ANS) processing, and describe the non-symbolic comparison tasks used to measure it. Pertinent issues relating the to measurement of the ANS are then discussed alongside implications for exploring its correlates, in particular mathematical achievement. Finally, I describe the aims and research questions addressed in the subsequent empirical chapters of this thesis.

1.1 Skills underlying mathematical competency

Competency in mathematics is a crucial skill required by most Western adults in everyday life. Mathematics is applied in multiple informal situations, for example, when calculating journey times, paying for goods in a shop, or when planning to re-decorate a room. Research has shown that poor attainment in school level numeracy is correlated with real-world practical difficulties such as defaulting on mortgage payments (Gerardi, Goette, & Meier, 2013) and poor budgeting for the future (Banks & Oldfield, 2007). In fact, greater mathematical competency has been found to lead to increased employability and higher salaries, over and above verbal skills (Parsons

& Bynner, 2005). Nonetheless, approximately 6%–14% of school-age children have persistent difficulties with mathematics despite age-appropriate achievement in other domains (Barbaresi, Katusic, Colligan, Weaver, & Jacobsen, 2005). Despite this high prevalence of mathematical difficulties, we do not yet have a thorough understanding of the skills that underlie achievement in this domain. Although in recent years there has been an increase in mathematical cognition research, evidence is piecemeal with many contradictory findings emerging in the literature. In order to provide suitable educational support for individuals with mathematics difficulties, it is necessary that researchers and educators continue to develop a more comprehensive understanding of the skills and underlying processes that are important for learning and performing mathematics.

Many of the cognitive skills that are already known to be involved in mathematics learning can be broadly categorised into two groups: domain-general skills and domain-specific skills. Domain-general skills include a broad range of processes that are not specifically related to mathematics, but could be applied to cognition any in domain, for example, working memory capacity. Domain-specific skills, on the other hand, include cognitive processes that are specifically related to mathematics, for example, counting knowledge. In the sections below I will provide further details and examples of both domain-general and domain-specific processing in relation to mathematics learning.

1.1.1 Domain-general skills

It is well established that formal mathematics abilities are substantially influenced by individual differences in domain-general processing skills, including inhibitory control, working memory, cognitive flexibility, and processing speed (Bull & Johnston, 1997; Bull, Johnston, & Roy, 1999; Bull, Espy, & Wiebe, 2008; Cragg & Gilmore, 2014; Friso-van den Bos, van der Ven, Kroesbergen, & van Luit, 2013; LeFevre et al., 2010). Executive functioning is the umbrella term for these cognitive processes that allow us to regulate behaviour in order to achieve goals, and respond flexibly to our changing environment (Cragg & Gilmore, 2014). Evidence to support the role of each executive function skill in mathematics learning and performance has been demonstrated empirically.

To begin, the influence of inhibitory control skills on mathematics achieve-

ment has received considerable research attention. Inhibition can be defined as an executive function mechanism that facilitates the suppression of prepotent responses in favour of efficient task processing (Dempster, 1992). Inhibition involves the ability to focus on task-relevant stimuli whilst resisting strong or automatic interference from task-irrelevant information. This is a skill often required not only at the level of calculation in mathematics, but also in terms of classroom behaviour more generally. For example, a child learning to determine the larger of two fractions, such as $\frac{1}{4}$ vs. $\frac{1}{8}$, would need to inhibit their previous knowledge of natural numbers (8 is bigger than 4), and focus on their newly acquired rational number knowledge (larger denominator = smaller fractional parts) to obtain the correct answer (Van Hoof, Janssen, Verschaffel, & Van Dooren, 2015). Similarly, in the classroom, children are required to use their inhibition skills on a more global scale to ignore distractions around the room, such as other children talking, to process the relevant information necessary to complete the work. Many different studies have found links between inhibitory control skills and mathematics ability (Blair & Razza, 2007; Bull & Scerif, 2001; Espy et al., 2004; St Clair-Thompson & Gathercole, 2001), and it is now widely accepted that individuals with better inhibition skills also tend to display higher performance on tasks measuring mathematical ability. The role of inhibition for both symbolic mathematics and non-symbolic estimation is discussed in greater detail in Part III of this thesis.

Another domain-general skill that has received a lot of attention from mathematical cognition researchers is working memory. Working memory refers to the temporary maintenance and manipulation of information required for complex processing (Baddeley, 1992), and is involved in tasks where information must be held in mind whilst new information is processed to obtain a solution. A superior working memory supports many of the mathematical procedures that involve multiple processing steps, for example, carrying numbers (DeStefano & LeFevre, 2004). Working memory has been found to be involved in children's basic mathematical processing (Bull & Espy, 2006), and also in more complex mathematical tasks such as multi-digit multiplication completed by adults (Tronsky, 2005). Accordingly, it is not surprising that researchers have found that individuals with a larger working memory capacity are more likely to score highly on mathematics achievement tests, in comparison to those with a smaller working mem-

ory capacity (Gathercole, Pickering, Knight, & Stegmann, 2004; McLean & Hitch, 1999).

The ability to think flexibly and shift fluently between closely related, yet distinct, conceptual representations is a critical skill for performing many academic tasks (Yeniad, Malda, Mesman, van IJzendoorn, & Pieper, 2013). In particular, this skill is thought to be important for mathematics processing due to the requirement to shift between different stages of a multi-step problem, from one arithmetic strategy (e.g. addition) to another (e.g. multiplication), or simply between verbal digits and Arabic symbols (Bull & Lee, 2014). Bull et al. (1999) found that children who performed less accurately on the Wisconsin Card Sorting Task, a task designed to measure cognitive flexibility and shifting, also demonstrated lower arithmetic test scores, even after controlling for reading attainment and IQ. This study found that the children with low mathematics abilities had particular difficulties with shifting from one sorting rule to another, a competency required for success on the varied range of skills measured by mathematics achievement tests (Bull et al., 1999). Several supporting studies have since replicated evidence of this relationship between shifting and mathematics achievement in a range of age groups and using a variety of shifting tasks (Andersson, 2010; Blair & Razza, 2007; Bull et al., 2008; see Yeniad et al., 2013 for a meta-analysis).

Alongside these complex executive function processes, very basic skills including attending to the task demands and the speed at which individuals process information have been shown to be good predictors of mathematics achievement. Unsurprisingly, attention is necessary to process information required to successfully complete a task. Attentive behaviour is most commonly measured using the Inattentive sub-scale of the Strengths and Weaknesses of ADHD Symptoms and Normal Behavior Scale (SWAN) (Swanson et al., 2001), which uses teacher ratings of classroom attentiveness across nine items. Studies using this measure have found that attentive behaviour independently predicts children's strategy development and performance on arithmetic word problems (Fuchs et al., 2010, 2013; Geary, Hoard, & Nugent, 2012). However, Fuchs et al. (2010) warns that this relationship would also emerge if teacher's judgements of attention are clouded by their knowledge of the child's academic ability. Therefore, teacher ratings may fail to provide an objective and reliable measure of attentive behaviour, instead serving as a proxy for achievement.

Finally, research suggests that individual differences in processing speed have a significant impact on mathematical competency. Processing speed is the speed at which individuals are able to fluently perform simple and repetitive cognitive tasks (Flanagan, Ortiz, & Alfonso, 2013). A study by Bull and Johnston (1997) reported that children who were slower at processing task information on a visual matching task demonstrated more difficulties automating the basic arithmetic facts needed for mathematical proficiency. Fast processing speed is thought to support efficient counting skills and assist in the creation of links between problems and answers (Fuchs et al., 2013).

In summary, multiple aspects of mathematical processing have been shown to be related to various different executive function skills. Although there is a great deal of evidence supporting these links, studies vary in the mathematical skills they assess, from counting to general standardised mathematics achievement, and in the measures of executive functioning they explore, from processing speed to working memory. Bull and Lee (2014) highlight the pitfalls of relying on a single measure to estimate the relationship between executive functions and mathematics performance, and recommend that confirmatory analytical techniques are implemented to verify that tasks are measuring the intended latent variable. Lee, Bull, and Ho (2013) have found that young children's executive function skills cannot be differentiated from each other until formal schooling begins, and continue to become more distinct into adolescence. Correspondingly, the same issue of differentiation of abilities can be applied to the study of mathematical processing. In studies that employ a standardised measure of mathematics achievement assessing a range of complex skills, it is not possible to determine which specific aspects of mathematics involve the executive function skill measured. Therefore, although the importance of executive function skills for overall mathematics learning and achievement is clear, more work in this area is needed to determine the mechanisms by which specific skills are related.

1.1.2 Domain-specific skills

In addition to the general processing abilities described above, there are many domain-specific skills considered to be central to learning and performing mathematics. One of the most obvious fundamental skills is counting ability. Understandably, many aspects of mathematics rely on one's

understanding of counting procedures, particularly in the early stages of arithmetic learning (Desoete, Ceulemans, Roeyers, & Huylebroeck, 2009). Gelman and Gallistel (1978) identified five key counting principles typically learned in preschool years, including understanding of the following: numbers have a fixed and stable order; the last number used when counting represents the cardinality of the set; the one-to-one correspondence principle stating that every item should be tagged once with a unique tag; the abstraction principle stating that any collection of objects can be counted; and finally the order-irrelevance principle stating that so long as all other counting principles are obeyed, objects may be tagged in any sequence. More recent research has shown that young children's understanding of these principles, assessed by their ability to detect counting rule violations, significantly relates to early mathematics achievement (LeFevre et al., 2006), demonstrating the importance of this basic skill.

Knowledge of number facts, such as fast retrieval of the number bonds to 10, is another key skill believed to underlie broader mathematics achievement. Arithmetic facts are thought to be stored in long-term memory, and accessed quickly using direct retrieval (Ashcraft & Battaglia, 1978), although the precise nature of this mechanism is the subject of debate (Baroody, 1994). Research has demonstrated that that poor number fact retrieval, as measured by performance on a speeded arithmetic recall task, is a defining feature of mathematics difficulties for primary school aged children, despite good reading ability (Jordan, Hanich, & Kaplan, 2003b). A similar study corroborated these results showing that primary-aged children who performed more accurately on simple arithmetic sums within three seconds (therefore retrieving the answer rather than calculating it from scratch) demonstrated superior performance on the Woodcock Johnson Mathematics Composite task, a standardised measure of applied problem solving and calculations (Jordan, Hanich, & Kaplan, 2003a). Further evidence for this link comes from supporting studies that have similarly demonstrated a significant relationship between arithmetic fact retrieval and wider mathematics performance (Geary, Hamson, & Hoard, 2000; Hanich, Jordan, Kaplan, & Dick, 2001).

A large body of research in mathematical cognition and education focuses on both procedural and conceptual understanding of mathematics in relation to successful learning. Procedural competence refers to the ability to solve

mathematical problems quickly and efficiently. Using an example referred to above, procedural competence in counting could be demonstrated through the ability to successfully recite a count list or accurately count an array of objects. Conceptual understanding, on the other hand, refers to knowledge of the underlying relationships and key principles within mathematics. Again, using counting as an example, conceptual understanding could be demonstrated through the ability to detect counting rule violations, as measured in LeFevre et al.'s (2006) study described above. This would show an understanding of the concepts that contribute to successful counting, whereas, in comparison, a procedural count task may be performed accurately by rote, without demonstrating any underlying knowledge of counting concepts. Both procedural skill and conceptual understanding are consistently found to underpin mathematical performance (Baroody, 2003), and there is considerable debate in the literature as to the relative importance of each and the relationship between them (Hiebert, 2013).

Finally, many psychologists consider accurate numerical representations to be crucial for mathematical success. The ability to approximate, compare and manipulate quantities, informally known as one's "number sense", is often considered a fundamental foundation for mathematical proficiency (Dehaene, 1997). The Approximate Number System (ANS) is the cognitive system thought to represent such approximations of quantity, and can be measured in both symbolic and non-symbolic tasks where participants are required to make 'more' or 'less' judgements about quantity. Studies with adults, children, and even very young infant participants have found that individuals with a more precise ANS also perform better on measures of mathematical proficiency (Halberda, Ly, Wilmer, Naiman, & Germine, 2012; Halberda, Mazzocco, & Feigenson, 2008; Libertus, Feigenson, & Halberda, 2011; Libertus, Odic, & Halberda, 2012; Mazzocco, Feigenson, & Halberda, 2011b, 2011a; Piazza et al., 2010; Piazza, Pica, Izard, Spelke, & Dehaene, 2013; Starr, Libertus, & Brannon, 2015). Nevertheless, several studies have found conflicting results and failed to find evidence of a correlation between ANS acuity and mathematics achievement in children (Holloway & Ansari, 2009; Sasanguie, De Smedt, Defever, & Reynvoet, 2011; Sasanguie, Van den Bussche, & Reynvoet, 2012; Sasanguie, Göbel, Moll, Smets, & Reynvoet, 2013), and in adults (Castronovo & Göbel, 2012; Inglis, Attridge, Batchelor, & Gilmore, 2011; Price, Palmer, Battista, & Ansari, 2012; see De Smedt,

Noël, Gilmore, & Ansari, 2013, for a review). Despite mixed evidence, the relationship between non-symbolic discrimination and formal mathematics ability has caught the attention of many psychologists and education researchers, and reports of significant correlations have been featured in many mainstream media outlets. If the significant correlational findings were to stem from a causal link between ANS acuity and mathematics achievement, this finding would have considerable implications for educational interventions focussed on training ANS acuity. However, it is first essential that we understand how individuals' ANS acuities are measured, the reliability and validity of these measurements, and whether there could be any other cognitive factors mediating the relationship with mathematics achievement. The remainder of this chapter will review the literature surrounding theories of ANS acuity, its measurement and its correlates.

1.2 The Approximate Number System (ANS)

One of the first formal discussions of an "approximate system" for representing numerical quantities was formulated by Stanislas Dehaene in his book "The number sense: How the mind creates mathematics", originally released in 1997. Dehaene described a universal system present in adults, children and even animals, that allows for the comparison, addition and subtraction of quantities without counting. Instead of using exact calculations, tasks that draw on the ANS are thought to be solved using approximate representations of quantity. Studies have shown that children aged just six months old can reliably discriminate between large sets of items that differ by a ratio of 0.5, for example 8 vs. 16 dots (Xu & Spelke, 2000). Similarly, many non-human primates and other animals can accurately contrast approximate quantities that differ in numerosity, so long as the ratio difference between the two sets is sufficiently different from 1 (Emmerton, 1998; Hauser, Tsao, Garcia, & Spelke, 2003). Given the lack of formal mathematical knowledge very young babies and animals have the capacity to obtain, the ANS is believed to be an innate system present from birth (Dehaene, 1997). As humans develop, ANS representations have been shown to become more precise, and adults have been found to reliably discriminate between quantities differing by up to a 0.9 ratio (Pica, Lemer, Izard, & Dehaene, 2004).

The examples above illustrate that ANS performance is dependent on the ratio between the quantities to be compared. This is one of the key theoretical features of the ANS: the further the ratio is from 1, the easier it is to distinguish numerosity differences. According to the standard model of the ANS (Barth, La Mont, Lipton, & Spelke, 2005; Dehaene, 1997), numerosity judgements follow the Weber-Fechner law. That is, when an individual sees n objects they form an ANS representation of the quantity. This representation is drawn from a normal distribution with mean n and standard deviation wn. Here w, or the 'Weber fraction', is a parameter which can be used to index the acuity of an individual's ANS. When asked to compare two numerosities, say n and m, it is the ratio of these two quantities and the value of w that predicts an individual's probability of success. This is because where the $n : m$ ratio is close to one, the distributions of possible n and m representations overlap to a greater extent, and so the probability of an individual generating incorrectly ordered representations is higher. Consequently, individuals are more likely to make an error comparing, for example, 29 vs. 30 items in comparison to 20 vs. 30 items.

ANS acuities are believed to vary between individuals and are consequently thought to influence task performance. According to the standard model, those with a more precise ANS (i.e. a smaller w) generate representations closer to the actual numerosity more often. This is thought to be reflected in superior performance on tasks used to measure ANS acuity (Barth et al., 2005; Dehaene, 1997), of which further detail is provided in the following section (1.3.1).

The standard model of the ANS (Barth et al., 2005; Dehaene, 1997), described above, proposes that the ratio difference between the two to-be-compared numerosities and the individual's ANS acuity are the only factors that influence ANS task performance. However, more recently there has been substantial debate around how measures of the ANS are influenced by additional factors, such as executive function skills, and the influence of visual cues in non-symbolic processing. These factors are the focus of this thesis and will be discussed in depth after details of tasks used to measure ANS acuity and its relationship with mathematics have been addressed.

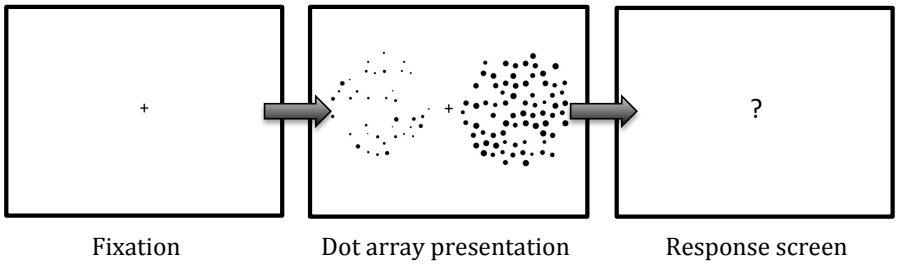


Figure 1.1: A typical dot comparison task trial procedure: First, participants view a central fixation point, followed by the brief presentation of two dot arrays, and finally a screen is presented to indicate that a response is required.

1.3 Measuring ANS acuity

1.3.1 Tasks

A range of different tasks have been developed to empirically measure an individual's ANS acuity. These include symbolic (e.g. digit) or non-symbolic (e.g. dot) approximate comparison and arithmetic tasks, estimation tasks, and even infant preferential-looking change detection paradigms (Barth et al., 2005; Gilmore, McCarthy, & Spelke, 2010; Starr et al., 2015; Xu & Spelke, 2000). Non-symbolic tasks may involve visual stimuli including arrays of dots or objects, auditory tone sequences, or a combination of these.

The most commonly-used and widely-accepted measure of the ANS is a dot comparison task (Price et al., 2012). This task involves the comparison two non-symbolic visual arrays of dots, across multiple different trials. (see Figure 1.1 for an example trial). During this task, the dot arrays are presented, usually for a very brief period of time to prohibit counting, and participants must estimate which array they believe has more dots in it. Participants can respond either by key press, verbally, or by pointing, depending on the presentation methods employed and the age of the participants. In order to achieve maximum control over stimuli presentation times, it is usual to present dot comparison tasks on a computer. Performance can be measured in terms of accuracy, response times, numerical ratio and distance effects or Weber fractions. Further discussion of the pros and cons of these measurements is provided later, in Section 1.3.3 of this literature review.

Tasks employing non-symbolic stimuli have also been developed to in-

vestigate the role of the ANS in approximate arithmetic, including addition and subtraction. The procedure of these tasks is more complex than simple comparison tasks (see Figure 1.2). Approximate arithmetic tasks usually begin with the presentation of a single initial array of dots. This array is subsequently covered by an occluder, usually a square shaped 'box'. During addition trials, a second array of dots moves in from the edge of the screen to behind the same occluder. In subtraction trials a second array moves out from behind the occluder and disappears off the edge of the screen. Finally, participants are required to compare the quantity of dots hypothetically remaining behind the occluder with a new comparison array, and respond based on which array is more numerous. As with dot comparison tasks, studies have shown that accuracy varies according to the ratio between the operation outcome and the comparison array, thus suggesting participants use their ANS representations to complete this task (Barth et al., 2005).

An alternative method used to measure ANS acuity is a non-symbolic numerosity estimation task. This task involves the display of individual arrays of dots and requires the participant to give a specific symbolic estimate of the number of dots presented in each array (Mejias, Grégoire, & Noël, 2012). The precision of the individual's estimate for this task is usually calculated by the absolute error score. Participants who make estimates with lower absolute error scores are thought to have more precise ANS representations.

Finally, the ANS acuity of infants, who are unable to respond to the above methods, has been assessed using preferential-looking paradigms. During such tasks, infants are presented with a series of non-symbolic dot arrays and the time spent looking at each array is measured. In one variation of a looking-time procedure used by Xu and Spelke (2000), a specific numerosity was presented repeatedly (with different patterns of dots) so that the infant habituated to this numerosity. Intermittently, a 'deviant' array representing a different numerosity was presented. The average time spent looking at this new array can be compared with the average looking time for the habituated array to gain a measure of ANS acuity (Xu & Spelke, 2000). In a similar procedure used by Starr et al. (2015), infants were shown two changing streams of numerosities presented side by side. One steam remained constant (e.g. 16, 16, 16) and one alternated between two numerosities (e.g. 8, 16, 8). Infants are thought to look longer at the changing stream if they are able to distinguish between the numerosities represented.

Figure 1.2: A typical non-symbolic addition trial animation procedure (adapted from Barth et al., 2005): Participants view an initial array of red dots, which is subsequently concealed by an occluder. Then a second array of red dots is added to the space behind the occluder. Lastly, participants are presented with a comparison array of blue dots and asked to decide whether there are more blue dots or red dots in total.

Although the use of Arabic numerals or non-symbolic dot arrays as stimuli is most prominent in the literature, tasks aiming to measure the ANS have also been successfully conducted using auditory stimuli. Barth et al. (2005) found that performance on both non-symbolic comparison and addition tasks was not influenced by the modality of the stimuli presented. Children aged between five and six years old completed the tasks either with dot array stimuli, or in a mixed format involving the substitution of one of the arrays with a sequence of sounds. Children were able to integrate quantity information from the two different modalities, and accuracy scores in the dual-modality task were not significantly different to scores in single visual modality task. However, due to ease of presentation, visually presented dot arrays are often the preferred choice of stimuli in the ANS literature.

In summary, a variety of tasks have been designed to measure the acuity of an individual's ANS representations. As aforementioned, dot comparison tasks are the most widely-used ANS task, and are considered the most direct measure of ANS acuity (Price et al., 2012). Price et al. (2012) suggest that assessing magnitudes in a symbolic format requires an additional processing step of mapping between symbols and magnitudes. Arithmetic tasks similarly require additional processing steps, over and above forming basic magnitude representations, to perform the addition or subtraction element of the task. Consequently, the remainder of this literature review will predominantly focus on comparison tasks with non-symbolic stimuli, rather than symbolic, arithmetic, estimation or mixed-modal tasks. This is to gain a clearer understanding of ANS processing distinct from additional cognitive processes such as mapping, or arithmetic.

1.3.2 Variations in dot comparison task methodologies

There is currently no universal procedure for conducting dot comparison tasks and consequently different studies have used diverse methods of presentation. The dot array stimuli can be presented in different formats, and can vary by the stimuli display times, the number of trials used, and the range of numerosities represented. Additionally, and importantly for this thesis, there is no consensus on how the visual characteristics of the dot arrays should be controlled. The problem with this lack of uniformity among dot comparison tasks used by different research groups is that we do not

know whether the same skills underlie performance on all variants of the task. There have been some attempts to disentangle the cognitive demands and the reliabilities of certain variations of dot comparison tasks (Price et al., 2012; Smets, Gebuis, Defever, & Reynvoet, 2014), focussed mainly on the format of the stimuli presentation.

One important variable that was inconsistent across research groups for some time was whether the comparison stimuli were presented sequentially (Ansari, Lyons, van Eimeren, & Xu, 2007), simultaneously side-by-side (Gilmore et al., 2013), or in an intermixed array with different coloured dots representing each set (Halberda et al., 2008) (see Figure 1.3 for an example of each). In 2012, Price and colleagues highlighted this lack of consistency within the literature and ran an experiment to establish the most reliable method of presentation. They found that the most robust method of stimuli presentation, in terms of reliability, is to display the dot arrays simultaneously, side-by-side on screen (Price et al., 2012). Price et al. reported that this method minimises the extraneous cognitive processing demands of the task. The sequential presentation of arrays is likely to involve increased working memory demands to hold and compare the numerosity information in mind once it has left the screen, and intermixed presentation of stimuli requires the additional visual processing demand of segregating visual information in order to make a comparison (Price et al., 2012). Since Price et al.'s publication, research groups that previously used intermixed designs (e.g. Halberda et al., 2008; Mazzocco et al., 2011a) have now begun using spatially separate simultaneous presentation methods (Libertus, Feigenson, & Halberda, 2013a, 2013b).

Differences in stimuli display times on dot comparison tasks can also substantially influence performance (Inglis & Gilmore, 2013). Inglis and Gilmore found that the longer an individual is given to process the stimuli on screen, the more precise the formation of the resultant ANS representation. Interestingly, participants were able to make above chance judgements about numerosity, even when stimuli were displayed for just 16 milliseconds, the refresh rate of the computer monitor. Nevertheless, when given 2400 milliseconds to view stimuli, participants performed significantly more accurately than at the lower presentation times. This finding implies that different processes may be recruited to complete ANS tasks in which the stimuli are presented very briefly, in comparison to ANS tasks where the

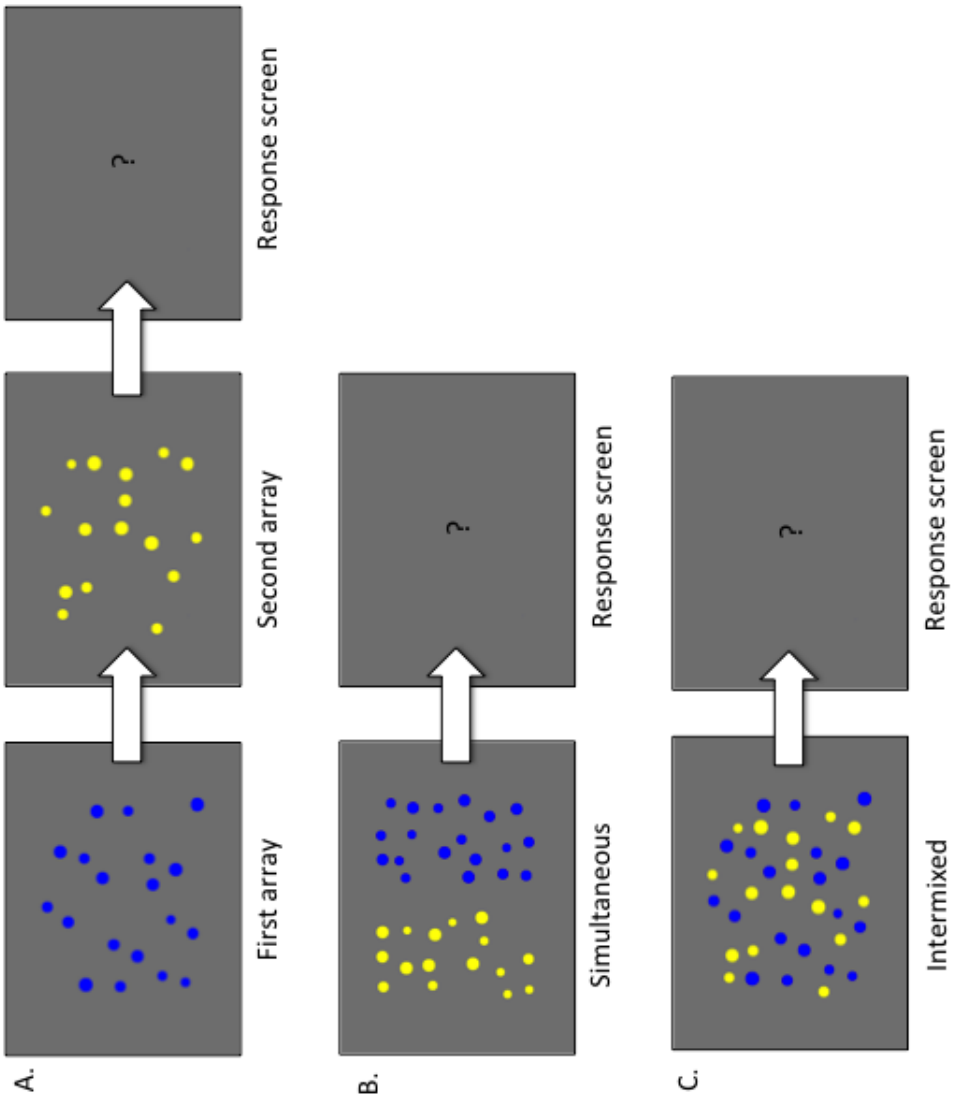


Figure 1.3: Examples of (a) sequential dot array presentation, (b) simultaneous presentation, and (c) intermixed presentation. Each trial usually begins with a fixation point in the centre of the screen.

participants are given longer to view the stimuli. Consequently, it is difficult to meaningfully compare findings from studies that use different stimuli presentation times. Furthermore, for studies where there is no fixed presentation time and participants are allowed to respond before the stimuli presentation duration is complete, there is likely to be a trade-off between accuracy and speed: those who respond faster may not perform as accurately as they would if they chose to view the array for the full length of the presentation. Therefore, it is recommended that tasks should employ fixed presentation times and use accuracy scores rather than response times for a more valid measure of the precision of ANS representations (Inglis & Gilmore, 2013).

The number of trials needed to provide a reliable and valid measure of ANS acuity is another unstandardised factor in the dot comparison task literature. In general, as the number of trials in a task increases, the reliability of the task also increases monotonically (Crocker & Algina, 1986). Non-symbolic comparison studies range from as few as 30 trials (Fuhs & McNeil, 2013), to as many as 750 trials (DeWind, Adams, Platt, & Brannon, 2015) in one task. In a recent review of the methodological differences in dot comparison tasks, Dietrich, Huber, and Nuerk (2015) summarised their recommendations for designing the most reliable and valid ANS task, and provided a checklist for doing so. The authors recommended, drawing from Lindskog and colleagues' work, that 400 trials are needed to reach an acceptable level of reliability (Lindskog, Winman, Juslin, & Poom, 2013). In their study, Lindskog et al. found that the split-half reliability of performance on dot comparison tasks with 50–200 trials is quite low (below 0.5), and only reaches an acceptable reliability of 0.7 at around 400 trials. In contrast, Gilmore, Attridge, and Inglis (2011), found that performance on a 120 trial dot comparison tasks had very good split-half reliability at 0.85 for small numerosity comparisons, and 0.96 for large numerosity comparisons. Similarly, Inglis and Gilmore (2013) found an acceptable immediate test-retest reliability of 0.68 for adults in just an 80 trial study. It is possible that these divergent reliability results stem from variation in the stimuli presentation methods. In Lindskog et al.'s (2013) study the stimuli were intermixed, and, as discussed above, Price et al. (2012) demonstrated that tasks using intermixed presentation methods were less reliable than tasks using spatially separate stimuli, presented simultaneously (as in Gilmore et al., 2011; Inglis

& Gilmore, 2013). Therefore, it is possible that in their checklist of recommendations, Dietrich et al. (2015) may have overestimated the number of trials necessary to obtain acceptable task reliability. Finally, Dietrich et al. (2015) recommended the use of an adaptive task procedure which takes account of performance on previous trials and adapts the difficulty level of forthcoming trials correspondingly. As stimuli are more diagnostic, i.e. only sampled from the region around the participant's accuracy threshold, Lindskog et al. (2013) have found this method to be a more economical way of gaining reliable dot comparison results with less trials. Nevertheless, this method means that it is not possible to use an average accuracy measure as the dependent variable because the difficulty of the trials is tailored to the individual's performance, i.e. participants will all end with similar average accuracy scores but may have completed different trials. Further research into the validity of this procedure is required.

Another factor that should be carefully considered in a numerosity processing task is the range of numerosities represented in the stimuli. However, the variable of stimuli set size does not appear to have been granted much attention in the ANS literature, and many studies use vastly different numerosity ranges. For example, Libertus et al. (2011) used a range of just 4–15 dots, whereas Inglis et al. (2011) used up to 70 dots in their stimuli. It is possible that the reason for this lack of standardisation within the literature stems from ANS theory. The dominant model of the ANS suggests that the absolute size of the to-be-compared numerosities should have no influence on accuracy scores for trials where the ratio is kept constant (Barth et al., 2005; Dehaene, 1997). For example, according to the this model, participants are equally likely to score correctly on a 7 vs. 10 dots trial as on a 70 vs. 100 dots trial. The only factors thought to influence dot comparison task performance are the ratios between the numerosities in each trial and the individual's ANS acuity (see section 1.2 for a discussion of this model). Following this theory, the effects of variation in absolute set size in dot comparison tasks are yet to be systematically explored. Some previous evidence suggests that variation of set size does not appear to influence task accuracy, but there are methodological limitations to these studies. Barth, Beckmann, and Spelke (2008) investigated whether set size affected dot comparison task accuracy by comparing the results from two of their studies that used different absolute set sizes. They found that set size had

no impact on accuracy scores, although the size of the sets explored only differed marginally between tasks (a 16 to 56 numerosity range, compared with a 5 to 40 range). Another study by Barth and colleagues reported corresponding results, but this study only examined response time performance and did not report accuracy scores (Barth, Kanwisher, & Spelke, 2002). It should be noted that small numerosities falling within the subitizing range (one to four) are thought to be processed using a different, more precise underlying cognitive mechanism to the ANS representations used for larger sets of items (Revkin, Piazza, Izard, Cohen, & Dehaene, 2008). In Dietrich et al.'s (2015) review of dot comparison task methodologies, they suggest that the subitizing range should be avoided, but do not give any specific advice concerning limits to the range of numerosities larger than four. Due to the limited knowledge regarding the effects of variation in the set size of stimuli on dot comparison performance, this is a topic that warrants further investigation and is discussed further in Chapter 5.

Finally, and importantly for the empirical work described subsequently in this thesis, there is no standard way of controlling the visual cues in non-symbolic comparison task stimuli. The dot arrays[1] are usually created with controls to ensure that the larger array does not always contain larger visual characteristics. This means that participants cannot rely on continuous visual cues alone to complete the task, and must make judgements based on numerosity discrimination to perform above chance level. There are multiple approaches to manipulating the visual characteristics of the stimuli, and different research groups tend to favour different approaches. The issue of visual cue control has come to be of increasing concern to researchers investigating the validity of dot comparison tasks, with some researchers suggesting that dot comparison tasks may be completed entirely through visual cue judgements (Gebuis & Reynvoet, 2012a). Therefore, the method of visual cue control is of fundamental importance and is explored further in Chapter 3 of this thesis.

The conclusion that numerous different methodological factors can influence dot comparison performance is problematic for the development of research into the ANS. As evidenced above, many studies have used diverse

[1]Dot arrays are the most commonly-used stimuli, but some studies have used groups of different shapes, e.g. stars (Fuhs & McNeil, 2013), or crayons (Mazzocco et al., 2011b).

dot comparison task methodologies, which renders it difficult to build on previous findings. The recent review of methodological variables by Dietrich et al. (2015) is valuable in highlighting these issues to the research community, and in its effort to provide a structure for a standardised task procedure. However, specific guidance for regulating some variables has still not been established and more work is needed to create a universal methodology that is a valid and reliable measure of the ANS.

1.3.3 Indexing ANS acuity

Alongside the many variations of ANS task methodologies, there are also a variety of dependent variables that are used to measure ANS acuity, including accuracy scores, response times, numerical ratio effects, numerical distance effects, and Weber fractions.

Accuracy scores are simply reported as the percentage of trials performed correctly across the entire task, or, for non-computerised tasks, the number of trials completed correctly within a time limit (e.g. Nosworthy, Bugden, Archibald, Evans, & Ansari, 2013). Individuals who demonstrate high accuracy scores on dot comparison tasks are thought to have more precise ANS representations. Similarly, with regard to response times, individuals who demonstrate faster responses on trials averaged across the task are also thought to have a more acute ANS (Halberda et al., 2012).

The numerical ratio effect (NRE) measures the influence of the numerosity ratios on task performance (Dietrich et al., 2015). As previously mentioned, the closer the ratio between the numerosities is to one, the slower and less accurate trial responses are likely to be. The NRE measures the level of increase in responses times or errors as the ratio between numerosities approaches one. The numerical distance effect (NDE) is a similar concept that measures the influence of the numerical distance between the arrays, rather than the ratio, on task performance (Dietrich et al., 2015). The NRE and NDE are indexed by calculating the size of the slope that relates either reaction time or accuracy to the numerical ratios (for NRE) and numerical distances (for NDE) between the to-be-compared arrays (Dietrich et al., 2015; Price et al., 2012). These two measures are highly correlated and often discussed interchangeably in the literature (Price et al., 2012). However, because NDE does not consider the absolute magnitude of the numerosities, NRE may better reflect ANS performance (Dietrich et al., 2015). A major

limitation of both measures as a valid index of ANS acuity is that a smaller (i.e. better) NRE or NDE could reflect floor effects rather than superior performance. A participant performing close to chance level would demonstrate similar performance on both easy and difficult ratio and numerical distance trials, and therefore would have a relatively flat regression slope, which could be wrongly interpreted as evidence of an acute ANS (Dietrich et al., 2015).

The Weber fraction (commonly referred to as w score) is a more complex measure of ANS acuity based on the assumption that dot comparison performance follows the Weber-Fechner law. As previously mentioned in section 1.2, Weber fractions represent the standard deviation or 'noisiness' of an individual's representation of magnitudes, with lower scores indicating higher precision. Weber fractions can be calculated from the following formula:

$$a = \frac{1}{2} + \frac{1}{2}\text{erf}\left(\frac{|n_1 - n_2|}{\sqrt{2}w\sqrt{n_1^2 + n_2^2}}\right)$$

Here, a represents the individual's overall accuracy, and n_1 and n_2 represent the to-be-compared numerosities. An individual's Weber fraction can be estimated by calculating the best fit of w in the equation (Inglis & Gilmore, 2014).

There have been some suggestions in the literature that Weber scores are superior to other measures of the ANS because they can be used to compare performance on dot comparison tasks that employ divergent methodological formats (Piazza et al., 2013). There is a commonly-held view that Weber fractions directly represent ANS acuity, independently from the numerical ratios and magnitudes used in the task (Dietrich et al., 2015). In contrast to this hypothesis, Inglis and Gilmore (2014) showed that both children's and adults' w scores are substantially influenced by the ratios in dot comparison task trials. Inglis and Gilmore demonstrated that there was a significant difference between the Weber scores of the same participants when calculated for easy ratios trials in comparison to more difficult ratios. Additionally, in a separate study described above (Section 1.3.2), Inglis and Gilmore showed that individuals' w scores also varied with the length of time the stimuli were displayed for, and therefore should not be used to compare performances between studies that use different display times (Inglis & Gilmore,

2013). Furthermore, Odic, Hock, and Halberda (2014) found that the order of trial presentation in a dot comparison task significantly affected w scores in a within-subjects design. Participants' w scores were superior on tasks that became increasingly more difficult, in comparison to tasks that became increasingly easier, despite both manipulations of the study containing exactly the same trials overall. These findings indicate that w scores are easily influenced by variants in task procedures, and consequently are not directly comparable between dot comparison task experiments using different designs. This has important implications for studies that have compared or combined participants' Weber fractions across different experiments (e.g. Halberda & Feigenson, 2008; Mazzocco et al., 2011b; Piazza et al., 2010), and indicate that conclusions from such studies may be flawed.

The dot comparison literature consists of studies reporting a variety of the measures described above, with the implicit assumption that they are all similar measures of ANS acuity. New studies build on findings from previous work with little consideration of the influence of these different measures. Inglis and Gilmore (2014) questioned this lack of consideration, and demonstrated the importance of assessing the psychometric properties and interrelations of ANS measures. In a study comparing participants' performance in terms of overall accuracy, NRE (accuracy), NRE (response time), and Weber Fractions, Inglis and Gilmore (2014) found that accuracy scores emerged as the best measure for dot comparison task performance. First, they found that NRE was a poor measure due to its low test-retest reliability after one week (<0.27), both for accuracy and response time NREs. Additionally, Inglis and Gilmore found that neither accuracy-based NREs or ratio-based NREs were significantly correlated with accuracy scores, Weber Fractions, or even each other. This provides strong evidence against the use of NREs to index ANS acuity. It was, however, found that w scores were highly correlated with accuracy scores ($R^2 = .86$), suggesting that these two measures may index the same cognitive construct. Nevertheless, in terms of psychometric properties, w scores were found to have a lower test-retest reliability than accuracy scores, and also follow a non-normal distribution. As accuracy scores followed a normal distribution and had the highest test-retest reliabilities for both adults and children, Inglis and Gilmore (2014) recommended that future researchers use accuracy figures rather than Weber fractions or NREs as a measure of dot comparison task performance.

The use of simple response time data as a preferred measure of ANS acuity can be ruled out by Inglis and Gilmore's (2013) work on the influence of display times on dot comparison performance, concluding instead that stimuli should be presented for a fixed duration (first discussed in section 1.3.2).

It could be argued that Weber fractions should be the preferred index of ANS acuity because they are a theoretically based measure derived from the well-established Weber-Fechner law. However, Weber fractions are contingent on the theory that dot comparison performance entirely follows the Weber-Fechner law, a claim that is yet to be validated. Considering that the cognitive underpinnings of dot comparison task performance are not yet fully understood, using a theoretically based measure of performance may be problematic. Therefore, a final justification for using accuracy over Weber fractions as an index of ANS acuity is that accuracy is an assumption-free measure of performance (Inglis & Gilmore, 2014).

Given the findings reported above, the dot comparison studies presented throughout the empirical chapters of this thesis all report accuracy as the dependent variable.

1.4 The relationship between ANS acuity and formal mathematics achievement

The ANS is claimed to be a basic cognitive system that we are born equipped with (Dehaene, 1997). The ANS is hypothesised to support approximate mathematical calculations such as estimated arithmetic, e.g. roughly 10 + roughly 20 = roughly 30, however is not precise enough to form representations necessary for exact calculations, e.g. 9 + 22 = 31. Conversely, formal symbolic mathematics is a learned skill, acquired through education. The link between informal ANS representations and formal mathematics achievement is one that has been greatly scrutinised by researchers (Chen & Li, 2014; De Smedt et al., 2013; Price et al., 2012). Feigenson, Dehaene, and Spelke (2004) propose that the ANS may represent a core system, or foundation, which supports more sophisticated higher-level mathematics.

If there happens to be a causal link between ANS acuity and mathematics ability, it follows that research may next focus on developing potential methods of refining ANS acuity in order to improve mathematics achievement. Evidence of this relationship could also assist in the identification of

students with mathematical difficulties, and similarly, identification of gifted mathematicians. However, though the literature provides some evidence of a correlation between the ANS and mathematics ability, results are mixed and often confounded, and the field is currently a long way from demonstrating evidence of a causal link. This section provides a brief review of the studies to date, and proposes some explanations for the disparate findings.

Throughout the last 10 years, many studies have reported a correlation between non-symbolic comparison performance and formal mathematics achievement in a the general population (see Table 1.1 for a summary and Chen & Li, 2014; Fazio, Bailey, Thompson, & Siegler, 2014 for meta-analyses). Of these studies, many have found there is a statistically significant relationship between an individual's ability to discriminate between two non-symbolic numerosities and their mathematical ability. That is, individuals who perform better on non-symbolic comparison tasks have also been shown to demonstrate better performance on tasks measuring formal, symbolic mathematics skills. Many of the high-profile studies reporting such a link in adults and children have been conducted by Halberda and colleagues (see Feigenson, Libertus, & Halberda, 2013, for a review). In an early influential paper published in *Nature*, Halberda et al. (2008) found that typically developing adolescents' Weber fraction scores, obtained at age 14, correlated with mathematics achievement data from the previous 10 years of schooling (as measured by the Test of Early Mathematical Ability, the TEMA-2, and the Woodcock-Johnson Calculation Subtest). The same research group later replicated this correlation between achievement on the TEMA-3 and dot comparison task performance with children as young as three years of age (Libertus et al., 2011).

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Agrillo, Piffer, and Adriano (2013)	Sequential	Adults	Acc	Mental arithmetic	$r = .463^{**}$
	Sequential	Adults	Acc	Mathematical Reasoning (WAIS-R)	$r = .489^{**}$
	Sequential	Adults	RT	Mental arithmetic	$r = .391^{*}$
	Sequential	Adults	RT	Mathematical Reasoning (WAIS-R)	$r = .449^{**}$
Bartelet, Vaessen, Blomert, and Ansari (2014)	Simultaneous	Children	RT	Arithmetic fact retrieval (TTA)	$r = -.14$
	Simultaneous	Children	Acc	Arithmetic fact retrieval (TTA)	$r = .24^{*}$

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Bonny and Lourenco (2013)	Simultaneous	Children	ANS precision (predicted for untested ratio)	TEMA-3	$r = .387^{***}$
Brankaer, Ghesquière, and De Smedt (2014)	Simultaneous	Children (6 years)	Acc	Tempo Test Arithmetic	$r = .36^{*}$
	Simultaneous	Children (6 years)	Acc	Curriculum-based standardised test	$r = .15$
	Simultaneous	Children (6 years)	RT	Tempo Test Arithmetic (TTA)	$r = -.13$
	Simultaneous	Children (6 years)	RT	Curriculum-based standardised test (untimed)	$r = .02$
	Simultaneous	Children (8 years)	Acc	Tempo Test Arithmetic (TTA)	$r = .14$

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
	Simultaneous	Children (8 years)	Acc	Curriculum-based standardised test	$r = -.16$
	Simultaneous	Children (8 years)	RT	Tempo Test Arithmetic (TTA)	$r = -.17$
	Simultaneous	Children (8 years)	RT	Curriculum-based standardised test	$r = -.20$
Fazio et al. (2014)	Simultaneous	Children	w and RT combined	School mathematics assessment (PSSA) score	$r = .60^{*}$
Fuhs and McNeil (2013)	Simultaneous	Children	Acc	TEMA-3	$r = .19$
Gilmore et al. (2013)	Simultaneous	Children	Acc	WJ-III Calculation subtest	$r = .57^{***}$

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Guillaume, Nys, Mussolin, and Content (2013)	Simultaneous	Adults	w	Addition arithmetic RT	$r = .47^{**}$
Halberda et al. (2008)[2]	Intermixed	Children (5 years)	w	TEMA-2	$r = .370^{**}$
	Intermixed	Children (5 years)	w	WJ-Rcalc	$r = .356^{**}$
	Intermixed	Children (6 years)	w	TEMA-2	$r = .374^{**}$
	Intermixed	Children (6 years)	w	WJ-Rcalc	$r = .571^{***}$
	Intermixed	Children (7 years)	w	TEMA-2	$r = .488^{***}$

[2]Dot comparison performance measured at 14 years, mathematics achievement measured at different time points provided in table.

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
	Intermixed	Children (8 years)	w	TEMA-2	$r = .569^{***}$
	Intermixed	Children (8 years)	w	WJ-Rcalc	$r = .531^{***}$
	Intermixed	Children (9 years)	w	WJ-Rcalc	$r = .498^{***}$
	Intermixed	Children (10 years)	w	WJ-Rcalc	$r = .342^{**}$
	Intermixed	Children (11 years)	w	WJ-Rcalc	$r = .501^{***}$
Halberda et al. (2012)	Intermixed	Children, Adults	w	Self-reported school mathematics achievement	$r = -.19^{***}$
	Intermixed	Children, Adults	RT	Self-reported school mathematics achievement	$r = -.09^{***}$

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Holloway and Ansari (2009)	Simultaneous	Children	NDE	WJ-III Mathematics Fluency and Calculation composite	$r = -.015$
Inglis et al. (2011)	Simultaneous	Children	w	WJ-III Calculation subtest	$r = -.548^{**}$[3]
	Simultaneous	Adults	w	WJ-III Calculation subtest	$r = .161$[3]
Kolkman, Kroesbergen, and Leseman (2013)	Simultaneous	Children	Acc	Standardised mathematics test	$r = .16$

[3]Partial correlation controlling for non-verbal IQ and age.

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Libertus et al. (2011)	Simultaneous	Children	Acc	TEMA-3	$r = -.424^{***}$
	Simultaneous	Children	w	TEMA-3	$r = -.265^{**}$
	Simultaneous	Children	RT	TEMA-3	$r = -.283^{***}$
Libertus et al. (2012)	Simultaneous	Adults	w	Scholastic Aptitude Test (SAT) Quantitative	$r = -.22^{*}$
Libertus et al. (2013a)	Simultaneous	Children	Acc	TEMA-3	$r = .52^{**}$
	Simultaneous	Children	w	TEMA-3	$r = -.42^{**}$
	Simultaneous	Children	RT	TEMA-3	$r = -.36^{**}$
Libertus et al. (2013b)	Simultaneous	Children	Acc	TEMA-3 informal mathematics items	$r = .44^{***}$
	Simultaneous	Children	Acc	TEMA-3 formal mathematics items	$r = .06$

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Lonnemann, Linkersdörfer, Hasselhorn, and Lindberg (2015)	Simultaneous	Children	NDE	Addition arithmetic	$r = -.04$
	Simultaneous	Children	NDE	Subtraction arithmetic	$r = .01$
Lourenco, Bonny, Fernandez, and Rao (2012)	Intermixed	Adults	Acc	WJ-III Calculation subtest	$r = .320^{**}$
	Intermixed	Adults	Acc	KeyMath 3 Geometry subtest	$r = .332^{***}$
Lyons, Price, Vaessen, Blomert, and Ansari (2014)	Simultaneous	Children	Acc and RT combined	Tempo Test Automatiseren (TTA)	$r = .554^{***}$

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Lyons and Beilock (2011)	Simultaneous	Adults	w	Mental arithmetic	$r = -.339^{*}$
Mazzocco et al. (2011b)[4]	Simultaneous	Children	Acc	TEMA-3	$r = -.527^{*}$
	Simultaneous	Children	w	TEMA-3	$r = -.456$
Mundy and Gilmore (2009)	Simultaneous	Children	Acc	Curriculum-based mathematics test	$r = .35$
	Simultaneous	Children	NDE	Curriculum-based mathematics test	$r = .02$
Nys and Content (2012)	Simultaneous	Adults	Acc	Tempo Test Rekenen (TTR)	$r = .16$

[4]Dot comparison performance measured at age 3–4 years (scores adjusted for age and display time at initial testing), TEMA-3 measured at 6-7 years (scores adjusted for age and grade at follow-up testing).

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
	Simultaneous	Adults	RT	Tempo Test Rekenen (TTR)	$r = -.08$
Price et al. (2012)	Sequential	Adults	NRE	WJ Math Fluency subtest	$r = .01$
	Simultaneous	Adults	NRE	WJ Math Fluency subtest	$r = .01$
	Intermixed	Adults	NRE	WJ Math Fluency subtest	$r = .03$
	Sequential	Adults	w	WJ Math Fluency subtest	$r = .10$
	Simultaneous	Adults	w	WJ Math Fluency subtest	$r = -.28$
	Intermixed	Adults	w	WJ Math Fluency subtest	$r = -.24$
Sasanguie et al. (2011)	Simultaneous	Children	RT/Error	Curriculum-based standardised test	$r = -.16$[5]
	Simultaneous	Children	NDE	Curriculum-based standardised test	$r = .08$[5]
Sasanguie et al. (2012)	Simultaneous	Children	RT/Error	Curriculum-based standardised test	$r = -.18$[5]

[5]Partial correlation controlling for grade (year group).

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
	Simultaneous	Children	NDE	Curriculum-based standardised test	$r = -.12$[5]
Sasanguie et al. (2013)	Simultaneous	Children	Acc	Tempo Test Rekenen (TTR)	$r = .14$[6]
	Simultaneous	Children	w	Tempo Test Rekenen (TTR)	$r = -.17$[6]
	Simultaneous	Children	Acc	Curriculum-based standardised test	$r = .09$[6]
	Simultaneous	Children	w	Curriculum-based standardised test	$r = -.17$[6]
Soto-Calvo, Simmons, Willis, and Adams (2015)	Simultaneous	Children	Acc	WIAT-II Mathematical Reasoning subtest	$r = .34^{***}$

[6]Partial correlation controlling for grade (year group) and spelling achievement.

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
	Simultaneous	Children	Acc	WIAT-II Numerical Operations subtest	$r = .39^{***}$
Starr et al. (2015)	Simultaneous	Children	w	TEMA-3	$r = -.42^{**}$
Vanbinst, Ghesquière, and De Smedt (2012)	Simultaneous	Children	NDE	Curriculum-based standardised test	$r = .03$
Zhou, Wei, Zhang, Cui, and Chen (2015)	Simultaneous	Children (8 years)	acc	School achievement test	$r = .28^{**}$
	Simultaneous	Children (8 years)	RT	School achievement test	$r = .24^{**}$
	Simultaneous	Children (9 years)	acc	School achievement test	$r = .18^{*}$
	Simultaneous	Children (9 years)	RT	School achievement test	$r = .03$

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
	Simultaneous	Children (10 years)	acc	School achievement test	$r = .25^{**}$
	Simultaneous	Children (10 years)	RT	School achievement test	$r = .06$

Table 1.1: A summary of the studies that have reported the relationship between non-symbolic comparison task performance and formal mathematics abilities in a typical population (both adults and children). The Pearson's correlation coefficients are provided, along with key characteristics of the studies including the stimuli presentation method, the age group of the participants, the index of non-symbolic comparison performance employed, and the mathematics ability measure. Acc = accuracy, RT = response time, w = Weber fraction, NDE = numerical distance effect, NRE= numerical ratio effect. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Likewise, several studies have found corresponding results with adult participants. For example, Libertus et al. (2012) demonstrated a significant correlation between adults' non-symbolic comparison task performance and their achievement on the quantitative section of the standardised college-entrance exams in the USA. Similarly, other studies have reported further consistent findings from participants across a wide range of age groups, from infancy (Starr et al., 2015) to older adults (Halberda et al., 2012).

In line with the above findings, research has shown that poor performance on non-symbolic comparison tasks can distinguish children with mathematical learning disabilities from their typically performing peers. Specifically, Mazzocco et al. (2011a) found that 14–15 year old students with dyscalculia demonstrated significantly lower w scores on a dot comparison task than their age-matched peers. Notably, this study found a significant difference between the dot comparison performance of students with dyscalculia and the performance of mathematically low-achieving students, but no significant difference between the low-achieving and typically-achieving students' performances. The authors suggested that this finding provides evidence that an ANS deficit may be specific to dyscalculia.

Further research in support of the relationship between dot comparison performance and formal mathematics ability comes from Piazza and colleagues' study of an Amazonian indigenous group, the Mundurucú (Piazza et al., 2013). This population have variable access to education, with availability determined by their proximity to the few schools in the area. Therefore, Piazza et al. (2013) were able to assess dot comparison task performance in adults and children with and without previous experience of formal schooling, from the same culture. They found that formal education was significantly associated with improved dot comparison task performance, independent of age. Specifically, this effect of education was only evident in participants who had reached the point in schooling where the curriculum began to involve learning arithmetic for the first time. From these results, it appears possible that access to mathematics education may improve the acuity with which individuals can represent non-symbolic quantities. These findings can be seen to bolster the above evidence provided from Western cultures, suggesting that there is a significant positive correlation between dot comparison task performance and mathematics ability that appears to be cross-cultural.

Nevertheless, in contrast to the evidence in support of a significant link between dot comparison performance and mathematics ability, there are conflicting findings from a number of studies that have failed to find this relationship (see Table 1.1, and De Smedt et al., 2013, for a review). As can be seen in Table 1.1 by the non-significnat correlation coefficients, studies demonstrating that dot comparison task performance did not significantly correlate with formal mathematical achievement are relatively common. Holloway and Ansari (2009), for example, demonstrated that children's performance on a non-symbolic comparison task was statistically unrelated to their mathematics fluency or calculation scores on the Woodcock Johnson III Test of Achievement. Similarly, in an adult population Price et al. (2012) found no significant correlation between arithmetic competency and performance on dot comparison tasks presented in three different formats (sequential, simultaneous and intermixed) and indexed by two different measures (w and NDE). Others have suggested that the relationship may depend upon developmental differences, and have reported a significant relationship between dot comparison performance and mathematics achievement in children, but not in adult participants (Inglis et al., 2011).

The summary of studies, presented in Table 1.1, that have investigated the relationship between dot comparison task performance and mathematics ability demonstrates not only the prominence of this research area from the large number of recent studies, but, importantly, the extent of the variability in results. Studies using corresponding age groups and the same index of measurements commonly report contrasting findings. In addition to the variables provided in Table 1.1, other factors are also likely to influence this relationship, including the sample size, and variations in dot comparison methodologies such as the stimuli presentation times, the number of trials presented, and the method by which the stimuli are created (see Section 2.3 for more detail on this).

Despite the emergence of these conflicting findings, the results from studies that have found a significant link between dot comparison performance and mathematics achievement have led to the hypothesis that it may be possible to improve mathematics ability by training ANS acuity using non-symbolic tasks (DeWind & Brannon, 2012; Hyde, Khanum, & Spelke, 2014; Park & Brannon, 2013). A study by Park and Brannon (2013) has demonstrated modest success from a training task that involved a non-symbolic

arithmetic task. Adult participants who practiced approximate addition and subtraction using dot arrays across 10 training sessions (see Figure 1.2, Section 1.3 for an example of an approximate arithmetic trial) significantly improved on a subsequent symbolic arithmetic task relative to a control group. Nevertheless, to date, there is no substantial evidence for improvements in formal mathematics following non-symbolic comparison training. Hyde et al. (2014) found that brief training on a dot comparison task (60 trials) led to improvements in 6–7 year old children's response times to arithmetic questions, but not accuracy. Interestingly, the children in this study showed no significant improvement in subsequent dot comparison task performance following the training, so it is unlikely that the decrease in response time to complete arithmetic questions was due to improved ANS acuity. Therefore, although attempts have been made to explore the causality of the relationship between non-symbolic comparison judgements and mathematics abilities through training studies, at present there is no convincing evidence to support this endeavour.

In line with the mixed evidence presented in Table 1.1 and the absence of successful ANS training studies, several researchers have highlighted that the link between ANS task performance and mathematics is poorly understood (e.g. Chen & Li, 2014; De Smedt et al., 2013; Price et al., 2012). It is therefore essential to develop a comprehensive understanding of the cognitive processes that underlie dot comparison performance before the potential educational applications of the task, such as training, are explored. Without a comprehensive knowledge of the cognitive skills that are involved in completing a dot comparison task, we cannot begin to make sense of correlations between performance on the dot comparison task and other cognitive abilities, including mathematics ability. We first need to understand whether variations in dot comparison methods influence the extent to which performance on the tasks reflects ANS acuity, and whether tasks are measuring more than just ANS acuity. Until a more complete picture of the factors that influence ANS task accuracy has been developed, it is premature to draw conclusions from correlations of dot comparison performance and mathematics achievement scores alone.

1.5 Summary

Research to date has demonstrated that a wide range of cognitive factors influence individuals' mathematics achievement, using cross-sectional and longitudinal methodologies with children and with adults. Evidence exists for multiple domain-general and domain-specific skills that contribute to mathematical success. Despite this wealth of evidence, the psychology of mathematical development is a large domain and still in its infancy. Evidence thus far is piecemeal and there remains a call for more research to uncover the complexities of individual skills, and how these skills interact with each other. This will aid the development of a comprehensive model of mathematical learning and achievement.

A notable gap in the literature surrounds the understanding of individual differences in ANS acuity and, specifically, the tasks used to measure it. For several years, dot comparison tasks have been presumed to be valid measures of ANS acuity. Research progressed very quickly from the development of the task—finding that performance was ratio dependent and therefore presumably measuring ANS acuity—to studies using dot comparison tasks to investigate the relationship between ANS representations and mathematics achievement. However, more recently, studies are beginning to emerge which investigate the basic psychometric properties of dot comparison tasks. These investigations suggest that the methodological variables in dot comparison tasks have a substantial impact on task performance. It follows that studies which employ different versions of the dot comparison task may not be measuring the ANS acuity to the same extent. This provides a threat to the validity of research that has built on previous studies investigating dot comparison task performance. Furthermore, we cannot be sure of the cognitive skills that each variant of the task requires, and how much ANS representations truly influence task performance. Before we continue to use dot comparison tasks to measure ANS acuity and its correlates, these issues must be addressed.

1.6 Research questions

The current thesis explores the cognitive skills that underlie dot comparison task performance. Specifically, the studies presented here focussed on the

visual characteristics of dot array stimuli, and the role of both domain-general processing and ANS acuity in dot comparison task performance.

The empirical findings are presented in three parts: Part II focusses on the influence of divergent methods of producing dot array stimuli, Part III reports on the role of inhibition in dot comparison tasks, and Part IV reports a re-analysis of the data which explores the relative influence of both visual cue and ANS processing on dot comparison performance. The research questions addressed in this thesis are as follows:

Part II: Visual cues in dot comparison tasks

Study 1. Do the visual cues in dot array stimuli influence task performance? Are tasks created with different controls for visual cues measuring the same cognitive construct? How reliable are these different methods?

Part III: Inhibition in dot comparison tasks

Study 2. How does the absolute set size, and the consequent change in the salience of visual characteristics in dot arrays, influence non-symbolic comparison task performance? Are responses in line with an inhibitory control account of performance?

Study 3. Does dot comparison task performance follow the same pattern of results as classic inhibition tasks?

Study 4 and Study 5. Does dot comparison task performance correlate with inhibition task performance?

Part IV: Do non-symbolic numerosity tasks involve numerosity processing?

Re-analysis of data. Do dot comparison tasks involve numerosity processing at all?

The results of these questions will be discussed in relation to the implications for the future of dot comparison tasks as measures of ANS acuity, and the relationship between ANS acuity and mathematics achievement.

Part II

Visual cues in dot comparison tasks

Chapter 2

Visual cues literature review

The main literature review presented in Chapter 1 introduced the subject of visual cue controls in dot array stimuli (Section 1.3.2). This section provides an in-depth review specifically focussed on the methods used to create dot array stimuli, and provides the details necessary for the empirical work presented later in this section.

2.1 Why control for visual characteristics in dot comparison stimuli?

The dot array stimuli presented in non-symbolic comparison tasks are usually produced with sophisticated computer-generated controls for continuous quantity variables which have the potential to bias responses to numerosity information. The visual characteristics of dot arrays, such as the size of the dots, are manipulated so that they are not consistently informative of number. If these variables were not systematically controlled, dot comparison tasks would involve a substantial confound: arrays with more numerous quantities would always contain larger visual properties. Consequently, it would not be possible to tell whether a participant had completed the task based on numerosity judgements, in accordance with the aims of the task, or whether they had simply responded based on visual property judgements.

To account for this confound, dot comparison tasks typically consist of both 'congruent' and 'incongruent' trials to control the relationship between the visual characteristics and the numerosity of the array. Congruent trials involve stimuli where the size of the visual characteristics of the arrays are

positively correlated with numerosity. Conversely, incongruent trials involve stimuli where the size of the visual characteristics are negatively correlated with numerosity. Dot comparison tasks typically contain a balance of congruent and incongruent trials to ensure that if a participant were basing responses purely on visual cues, in conflict with task requirements, their overall accuracy score would not be significantly above chance level.

2.2 The visual characteristics of dot arrays

Although most studies report the need to control the relationship between numerosity and visual characteristics, the variables that researchers choose to control differ across studies. Gebuis and Reynvoet (2011) highlighted five visual characteristics that covary with numerosity, and can be manipulated in dot comparison task stimuli:

1. Convex-hull size: The smallest contour surrounding all of the dots in the array. This is sometimes known as the area extended, envelope area, or the occupied area of the dot array. See Figure 2.1 for an illustration.

2. Average dot size: The average diameter of the dots within the array, sometimes referred to as item size.

3. Total circumference: The cumulative circumference of all of the dots in one array, also referred to as contour length.

4. Cumulative surface area: The total surface area of all of the dot surfaces within the array. This can be referred to as total or aggregate surface area.

5. Density: The convex-hull size divided by the cumulative surface area.

Gebuis and Reynvoet (2012a) note that although there are five distinct visual aspects of dot arrays that can be measured, some of these aspects are highly correlated with each other. Gebuis and Reynvoet report that if cumulative surface area increases, the average dot size and density in the array also increases, whereas convex hull can remain constant. For this reason, in Gebuis and colleagues' papers (Defever, Reynvoet, & Gebuis, 2013; Gebuis & Reynvoet, 2012a, 2012b; Szűcs, Nobes, Devine, Gabriel, & Gebuis,

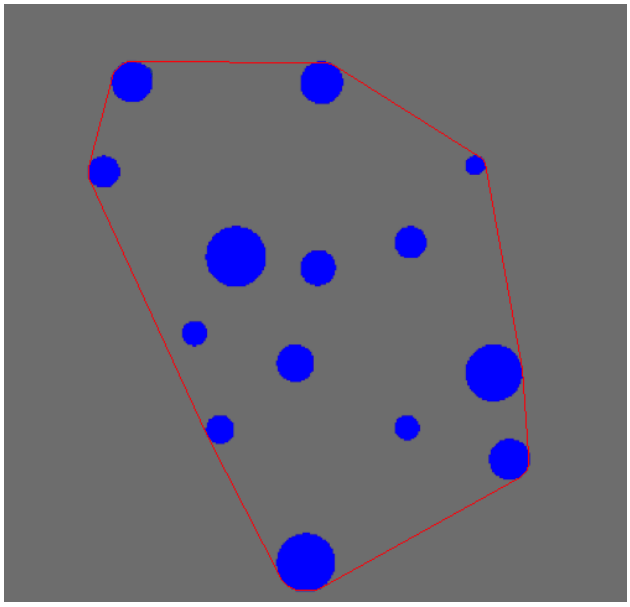


Figure 2.1: Convex hull, the smallest contour surrounding all of the dots in the array, is represented by the red line.

2015), and in the studies reported in this thesis, analyses exploring cumulative surface area, average dot size and density are combined as one factor. Besides Gebuis and Reynvoet's (2011) paper, the total circumference of dot arrays tends to be referenced less often than the other visual cues. Nevertheless, Szűcs et al. (2015) found the total circumference of dots in an array correlated highly with the other 'dot size' group of visual cues (cumulative surface area, dot size, density), and included it amongst this group. To conclude, there are multiple visual cues that can be referred to and analysed in dot array stimuli, however these fall into two categories, a 'dot size' category (cumulative surface area, dot size, density, total circumference), and a 'convex hull' category (convex hull). As the individual influence of the separate components in the dot size category cannot be disentangled, there is no benefit from reporting these visual cue variables separately in analyses (Gebuis & Reynvoet, 2012a, 2012b). A recent study by DeWind and colleagues supported this view and stated that a measure of 'size' (including cumulative surface area and individual dot size) and a measure of 'spacing' (including convex hull and sparsity) are together sufficient to determine the full set of features of dot array stimuli (DeWind et al., 2015).

It is important to note that there are several different variations of the terms used for each visual aspect of dot array stimuli. Given the lack of consistency in terms used to define both convex hull (area extended, envelope area, occupied area), and cumulative surface area (total surface area, aggregate surface area), the use of less specific terms, such as 'total area' (e.g. Halberda & Feigenson, 2008; Lindskog et al., 2013; Nys & Content, 2012; Odic, Pietroski, Hunter, Lidz, & Halberda, 2013), could create confusion between definitions. Visual cue terms are used interchangeably within the literature, and consequently it can be difficult to establish which visual characteristics have been considered in a study, and to compare this across different studies.

2.3 Methods of controlling for visual cues

Research groups vary in their approach to creating dot comparison stimuli. Different methods do not always control the same visual characteristics to the same extent. One frequently-used method of controlling the relationship between visual cues and numerosity in dot comparison stimuli is to manipulate the cumulative surface area and the average size of the dots. This is done by creating 50% of the task trials with stimulus pairs that have equal average dot size (the larger set has a larger cumulative surface area), and 50% of the trials with stimulus pairs that have equal cumulative surface area (the larger set has smaller average dot size) (see Figure 2.2 for an example). Libertus et al. (2012) refers to these trials as "correlated" and "anti-correlated", respectively, in terms of the relationship between cumulative surface area and numerosity. This method was first developed by Dehaene, Izard, and Piazza (2005) (Matlab script available at www.unicog.org/docs/DocumentationDotsGeneration.doc) and is thought to discourage the reliance on visual cues because no single cue is predictive throughout the entire task. This principle of visual cue control is also the default setting on the freely available Panamath software (Halberda et al., 2008; www.panamath.org), and has been used in multiple studies of the ANS (Halberda et al., 2008; Halberda & Feigenson, 2008; Halberda et al., 2012; Hellgren, Halberda, Forsman, Ådén, & Libertus, 2013; Libertus et al., 2011, 2012, 2013a, 2013b; Mazzocco et al., 2011a, 2011b; Odic et al., 2014; Odic, Libertus, Feigenson, & Halberda, 2013). Using this method, the

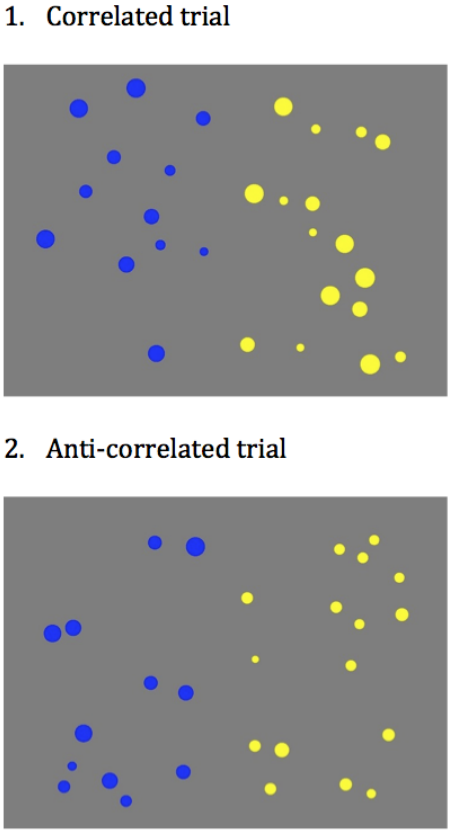


Figure 2.2: An example of a "correlated" (above) and "anti-correlated" (below) trial created with the Panamath protocol. The trial names refer to the relationship between the cumulative surface area and numerosity in each of the trials. Both stimuli represent a 12 vs. 16 dot trial.

convex hull of the dot arrays is not explicitly controlled.

In addition to the manipulation of cumulative surface area, Pica et al. (2004) developed a method that also controlled for the convex-hull size of the array. This method created 50% of trials where the larger numerosity contained a larger cumulative surface area and a larger convex hull, and 50% of the trials where the larger numerosity contained a smaller cumulative surface area and a smaller convex hull (see Figure 2.3). In this way, both visual cues varied together, either congruently or incongruently with the numerosity the array represented.

Gebuis and Reynvoet (2011) developed Pica et al.'s, (2004) protocol fur-

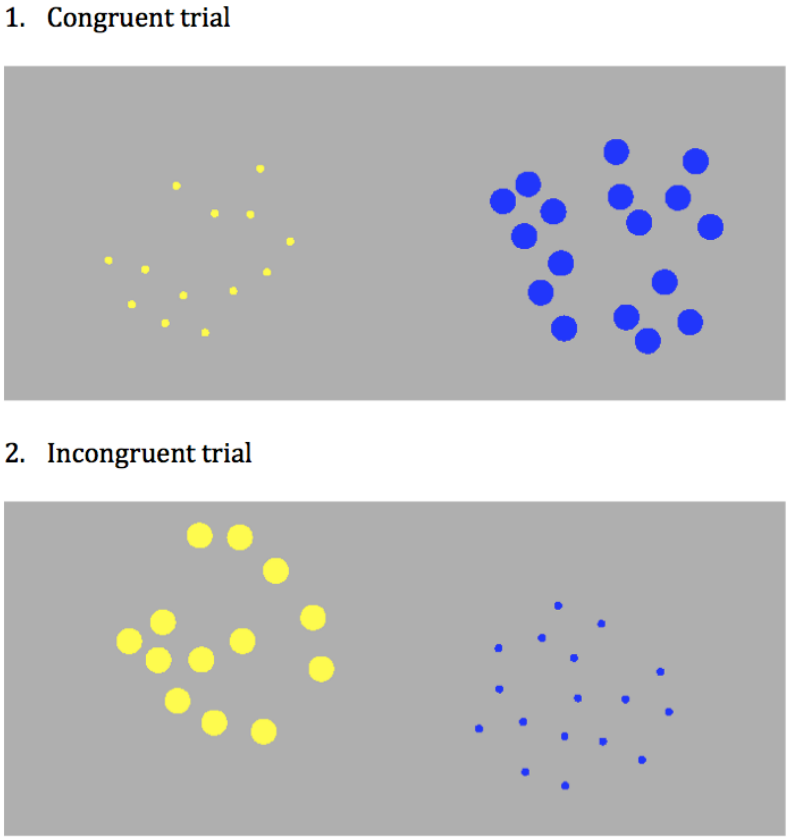
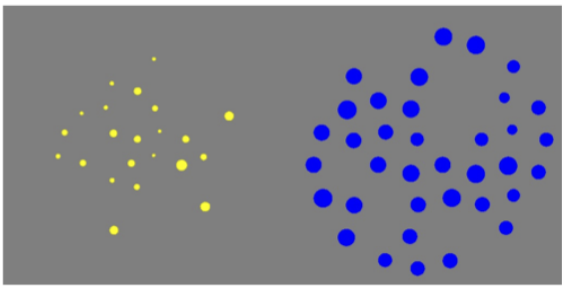


Figure 2.3: An example of a "congruent" and "incongruent" trial created with the Pica protocol. The trial names refer to the relationship between both cumulative surface area and convex hull with numerosity in each of the trials. Both stimuli represent a 13 vs. 17 dot trial.

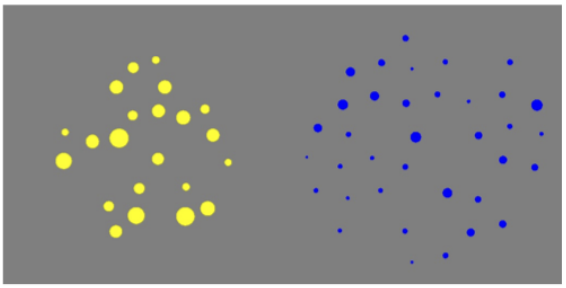
ther to create stimuli whereby both cumulative surface area and convex hull are accounted for. Using this protocol, trials can be partially congruent so that one visual cue is correlated with numerosity whilst the other is not. To elucidate, this method creates the following trials (see Figure 2.4 for example images): 25% of trials where the more numerous array has a larger cumulative surface area and a larger convex hull than its comparison array; 25% of trials where the more numerous array has a smaller cumulative surface area but a larger convex hull; 25% of trials where the more numerous array has a larger cumulative surface area but a smaller convex hull; and 25% of trials where the more numerous array has a smaller cumulative surface area and a smaller convex hull. Importantly, no single visual cue is consistently informative of numerosity in this method. Gebuis and Reynvoet (2011) criticised methods that only control for a single visual property at a time, and suggest that participants are likely to rely on multiple visual cues and switch between them depending on the trial characteristics. Gebuis and Reynvoet (2011) provide an example of a trial where one visual cue, e.g. average dot size, is equated across the two arrays and therefore uninformative of numerosity. Gebuis and Reynvoet suggest that in such a case, participants are likely switch their focus to an uncontrolled visual cue, e.g. cumulative surface area, which covaries with numerosity. Indeed, in a subsequent study, Gebuis and Reynvoet (2012a) showed that the influence of trial congruency on participants' dot comparison judgements increased when the number of visual cues controlled for increased. Participants showed smaller congruency effects (i.e. less difference in accuracy between congruent and incongruent trials) when only one visual cue was manipulated at a time than when multiple visual cues were manipulated. This suggests that participants are actually weighing up a range of non-numerical visual cues to help them make numerical judgements. Consequently, in recent years, researchers have begun to adopt Gebuis and Reynvoet's (2011) more comprehensive method of multiple visual cue control (Defever et al., 2013; Gebuis & Reynvoet, 2012a, 2012b; Gilmore et al., 2013; Inglis & Gilmore, 2013, 2014; Smets et al., 2014; Szűcs et al., 2015). Gebuis and Reynvoet's Matlab script to create dot arrays in this way is freely available online for other researchers to download (http://titiagebuis.eu/Materials.html).

It must be noted that Gebuis and Reynvoet (2011) warned that with small numerosities it is not always possible to control the visual cues in dot

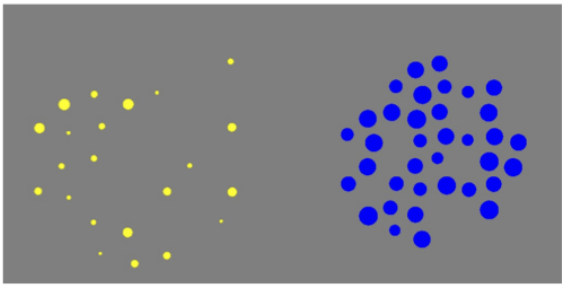
1. Fully congruent



2. Cumulative surface area incongruent, convex hull congruent



3. Cumulative surface area congruent, convex hull incongruent



4. Fully incongruent

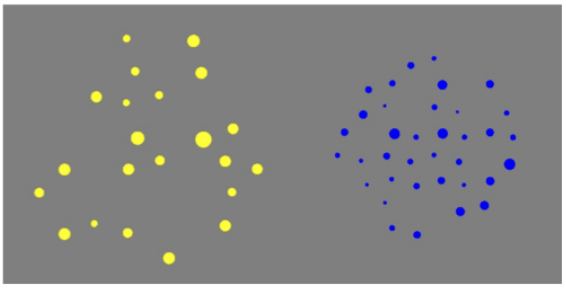


Figure 2.4: An example of the four image types created with the Gebuis and Reynvoet script. All images represent a 22 vs. 36 dot trial.

arrays as intended by their program. This is because the random placement of a small number of dots will not always spread to create the convex-hull size as desired. Gebuis and Reynvoet (2011) recommended registering the size of each visual cue for post hoc analyses to ensure there is no correlation between the visual cues and numerosity ratios across all trials. In addition to checking that the visual cues have been controlled correctly, Szűcs et al. (2015) recommended that researchers investigate how the relationship between particular visual cues and numerosity ratios influences participants' performances on the task. Szűcs et al. suggested that because it is physically impossible to create a single trial with visual properties that are 'truly neutral', i.e. there are always visual cues that correlate with number in any one particular trial, researchers should investigate how trials with different visual controls affect judgements, as well as looking at performance averaged across the whole task.

2.4 The visual cue account of dot comparison task performance

As mentioned previously, Gebuis and Reynvoet (2012a) found that variation in the number of visual cues controlled for in dot array stimuli influenced participants' dot comparison task accuracy scores. This has important implications for the comparison of studies in the literature that use different methods to create their stimuli. Gebuis and Reynvoet (2012a) concluded that less stringent designs may not be sufficient to ensure participants are not relying on the visual characteristics of the task to make their judgements, due to the finding that participants integrate multiple visual cues from the stimuli. Consequently, the level at which different dot comparison tasks are tapping the ANS, and how much visual processing can account for performance is unclear, and may vary from task to task.

In fact, Gebuis and Reynvoet proposed that the existence of an ANS that can extract quantity information independently from a visual scene appears unlikely (Gebuis & Reynvoet, 2012a). Instead, they proposed that when a participant is faced with an individual dot comparison trial, accuracy is influenced by their ability to attend to and 'weigh up' combinations of visual cues to make their choice (Gebuis & Reynvoet, 2012a, 2012b). This hypothesis contradicts the dominant standard model of the ANS which proposes

that approximate numerical judgements of quantity are made independently from the visual characteristics of the stimuli (Feigenson et al., 2004). Nevertheless, Gebuis and Reynvoet (2012a) state that more evidence is needed to fully support their theory.

2.5 The influence of different methods of visual cue control

Like many other variations in dot comparison task methodologies (described in Chapter 1, Section 1.3.2), the influence of disparities in visual cue control methods between studies has largely been ignored. Researchers have implicitly assumed that dot comparison tasks with differences in the way non-numerical cues are controlled provide equivalent measures of ANS acuity.

Recently, a study by Smets, Sasanguie, Szűcs, and Reynvoet (2015) investigated whether different methods for constructing dot array stimuli influenced adult participants' performance on both a numerosity estimation task and a dot comparison task. The stimuli construction methods contrasted in the study were the single visual cue control method developed by Dehaene et al. (2005), and the multiple visual cue control method developed by Gebuis and Reynvoet (2011). Participants completed identical numerosity trials in one testing session, created using these two divergent methods. If dot comparison tasks that employ different methods of visual cue controls are comparable measures of ANS acuity, one would expect that visual cue controls would not substantially influence task performance. In contrast to this, Smets et al. (2015) found a significant difference in participants' accuracy scores on each set of trials created with the two methods. The authors reported a non-significant correlation between participants' accuracy on the single cue condition and the multiple cue condition ($r = .23$). This is strong evidence to suggest that dot comparison judgements do not provide pure measures of ANS acuity that are independent from the visual characteristics of the stimuli.

Interestingly, Smets et al. (2015) found that differences in the protocol for controlling visual cues had no significant influence on participants performance on the numerosity estimation task, where participants are asked to estimate how many dots are in a single array. This finding implies that non-

symbolic numerosity estimation tasks involve different cognitive processes to the non-symbolic comparison task. Smets and colleagues hypothesised that the comparison task may encourage reliance on visual cues due to the simultaneous presentation of the stimuli and the requirement to simply select the more numerous array, rather than give an absolute estimate of numerosity (Smets et al., 2015).

2.6 Summary

For dot comparison tasks to be useful as measures of the ANS, a system believed to be able to extract numerosity information independently from visual cues (Feigenson et al., 2004), it is crucial that the visual characteristics in dot comparison stimuli are not informative of numerosity. There are multiple visual characteristics that can be measured in dot arrays, but many of these are highly correlated with each other so visual cues can be broadly categorised into two groups: dot size variables (including average dot size, cumulative surface area, total circumference and density), and convex hull. Researchers have developed different protocols for creating dot array stimuli, ranging from the control of individual visual cues (Dehaene et al., 2005; Halberda et al., 2008), to the manipulation of multiple visual cues simultaneously (Gebuis & Reynvoet, 2011). Therefore, visual cue controls are unstandardised across dot comparison studies within the literature, and the substantial influence of this factor on performance has only recently been demonstrated (Smets et al., 2015).

Chapter 3

Dot comparison stimuli are not all alike: The effect of different visual controls on ANS measurement (Study 1)

The following empirical study investigated how the accuracy and reliability of numerosity magnitude judgements are influenced by the visual controls in the stimuli. Although a similar study was conducted concurrently by Smets et al. (2015) (discussed above in Chapter 2, Section 2.5), the present study also investigated differences in test-retest reliability between methods, and how different visual cue controls influence performance congruency effects.[1]

3.1 Introduction

As discussed in Section 1.4 of the literature review, the link between ANS acuity and mathematics achievement has been widely debated, and conflicting results have been reported (see De Smedt et al., 2013 for a review). Studies investigating this link often assume dot comparison tasks provide valid and reliable measure of ANS acuity, but previous research has given only limited attention to the development of these tasks. The mixed findings

[1]The study presented in this chapter is published in *Acta Psychologica* (Clayton, Gilmore, & Inglis, 2015)

regarding the relationship between ANS task performance and mathematics ability could be, at least in part, due to the many differences in dot comparison task methodologies within the literature. Currently there is no standard protocol for creating dot array stimuli and it is unclear whether tasks that control for different visual cues measure the same cognitive constructs.

The review presented in Chapter 2, Section 2.3 describes the common methods used to control the visual characteristics of dot array stimuli so that they are not informative of numerosity across the entire task. This is intended to ensure that participants cannot perform significantly above chance simply by relying on non-numerical cues. A result of this manipulation is that half of the trials are congruent in terms of the relationship between numerosity and a particular visual cue size, and the other half are incongruent. Crucially, some studies have shown that participants perform more accurately on congruent trials, where the more numerous array also has larger visual characteristics, than incongruent trials, where the less numerous array has larger visual characteristics (Barth et al., 2006; Cappelletti, Didino, Stoianov, & Zorzi, 2014; Gilmore et al., 2013; Hurewitz, Gelman, & Schnitzer, 2006; Nys & Content, 2012; Szűcs et al., 2015). However, other studies have failed to find this effect (Gebuis & van der Smagt, 2011; Odic, Libertus, et al., 2013; Odic et al., 2014). It is possible that mixed results are partly due to divergent methodologies for controlling visual cues employed in the tasks. Notably, the studies which have not found corresponding congruency effects did not explicitly manipulate convex-hull size (Gebuis & van der Smagt, 2011; Odic, Libertus, et al., 2013; Odic et al., 2014). It is therefore important to understand more about when congruency effects occur, through a controlled comparison of different types of congruency in one group of participants.

The reliability of dot comparison tasks has been found to vary between tasks with different methodological formats (see Chapter 1, Section 1.3.2). For example, Price et al. (2012) found dot comparison tasks that used simultaneous presentation of dot arrays were significantly more reliable than those using intermixed or sequential stimuli. Other studies investigating the reliability of dot comparison tasks have focussed on the number of trials required to obtain a reliable measure of performance (Gilmore et al., 2011; Lindskog et al., 2013). Importantly, given recent evidence that visual cue controls substantially influence overall accuracy (Smets et al., 2015), it has

not yet been investigated whether differences in the visual characteristics of dot arrays stimuli influence the reliability of the task.

The present study aimed to investigate the test-retest reliability and concurrent validity of dot comparison tasks created using two different stimuli protocols. The first method, based on Dehaene et al.'s (2005) method of controlling visual cues, is the Panamath protocol (Halberda et al., 2008). This protocol been widely used in ANS research (e.g. Halberda et al., 2008; Halberda & Feigenson, 2008; Halberda et al., 2012; Hellgren et al., 2013; Libertus et al., 2011, 2012, 2013a, 2013b; Mazzocco et al., 2011a, 2011b; Odic et al., 2014; Odic, Libertus, et al., 2013) and manipulates, one at a time, either the average dot size or the cumulative surface area of the arrays. The second method used to create dot comparison stimuli is the Gebuis and Reynvoet (2011) protocol which controls for both cumulative surface area and convex-hull size simultaneously. This is also a commonly-used method of creating non-symbolic stimuli in research (e.g. Defever et al., 2013; Gebuis & Reynvoet, 2012a, 2012b; Gilmore et al., 2013; Inglis & Gilmore, 2013, 2014; Smets et al., 2014; Szűcs et al., 2015).

This study aimed to address three main research questions. First, is there a significant correlation between participants' accuracy scores on dot comparison trials created with the Panamath protocol and trials created with the Gebuis and Reynvoet protocol? Second, are there significant differences in the immediate test-retest reliabilities of each measure? Finally, do participants show congruency effects on trials created with both protocols? The answers to these questions will help to inform future research about the comparability of different protocols used to create stimuli to investigate ANS acuity and may provide explanations for conflicting evidence in the existing literature.

3.2 Method

3.2.1 Participants

Participants were 57 adult students[2] from Loughborough University (24 male, 33 female) with a mean age of 21.34 years (SD= 2.35). Participants were tested individually in a quiet room and were given a £3 inconvenience allowance for their time. This study was approved by the Loughborough University Ethics Approvals (Human Participants) Sub-Committee.

3.2.2 Task

Participants completed a nonsymbolic dot comparison task on a computer.[3] On each trial they were required to select the more numerous of two dot arrays. The two arrays consisted of blue or yellow dots on a grey background and were presented simultaneously, side-by-side on a 15" laptop screen. Participants were asked to select which array was more numerous using left and right keys marked on the keyboard. There were two types of dot comparison stimuli: arrays created using the Gebuis and Reynvoet (2011) protocol, and arrays created using Panamath software (www.panamath.org, Halberda et al., 2008), described in further detail in the Stimuli section below.

Participants completed eight practice trials followed by a total of 312 experimental trials, which were divided into four blocks. Block one consisted of 96 trials created with the Gebuis and Reynvoet (2011) protocol and block two consisted of 60 trials created with the Panamath protocol.[4] Both blocks were then repeated so that participants completed each trial twice in order to gain a measure of reliability. The order of blocks was counterbalanced so that half the participants completed block one first, and half completed block two first. Trials within the blocks were presented in a random order.

[2]The predictions for the results of this study, and all the studies presented in this thesis, apply equally to both adults and and children. Therefore, the choice of population used was based on pragmatic factors such as the length and difficulty of the task, and access to participants.

[3]Participants also completed an inhibitory control task in the same testing session, the results of which are reported in Chapter 7, Section 7.2.

[4]Different numbers of trials were included for each protocol as it was not possible to create sets of numerically matching trials using these two stimuli generation methods. Therefore trials were chosen to reflect the default use of each protocol in the literature. See section 3.2.3 for an in-depth discussion of this.

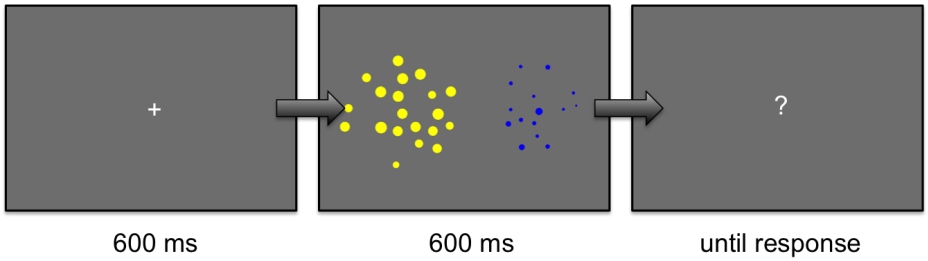


Figure 3.1: The dot comparison task trial procedure. Each trial began with a fixation point displayed for 600 ms, followed by the presentation of the dot arrays for 600 ms, and finally a question mark screen presented until the participant gave a response.

Each trial began with a fixation point (600 ms) followed the by presentation of the two arrays (600 ms) and finally a grey screen with a white '?' was presented in the centre until a response was given (see Figure 3.1). The task took approximately 15 minutes to complete.

3.2.3 Stimuli

The Panamath protocol stimuli were downloaded from an example of a pre-existing experiment available for research use on the Panamath website (http://www.panamath.org/9-12CollegeMaterials.zip; stimuli used in Libertus et al., 2012). Panamath stimuli can be classified as "correlated" and "anti-correlated" in terms of the cumulative surface area of the dots[5] and numerosity. Correlated trials included pairs of arrays where the more numerous array contained a larger cumulative surface area. Anti-correlated trials included pairs where the more numerous array contained a smaller cumulative surface area (see Figure 2.2 presented in Chapter 2, Section 2.3, for an example of Panamath stimuli). The colours of the dot arrays randomly alternated between blue and yellow on the left and right hand side of the screen.

The Gebuis and Reynvoet (2011) protocol stimuli were generated using a freely available Matlab script provided online (version May 20th 2011,

[5]For the stimuli used in this study, there was a high correlation between cumulative surface area and average dot size (r = .95) and density (r = .84). Consequently, for the remainder of the paper, only cumulative surface area is used in the analyses. The justifications for this are discussed in more detail in Chapter 2.3, Section 2.2.

http://titiagebuis.eu/Materials.html). This script controlled for cumulative surface area and convex hull, and generated four image types per trial (see Figure 2.4 presented in Chapter 2, Section 2.3). The first image type (fully congruent), included pairs of arrays where the more numerous array had a larger cumulative surface area and a larger convex hull. The second image type (cumulative surface area incongruent, convex-hull congruent), included pairs of arrays where the more numerous array had a smaller cumulative surface area and larger convex hull. The third image type (cumulative surface area congruent, convex-hull incongruent), included pairs of arrays where the more numerous array had a larger cumulative surface area and a smaller convex hull. The fourth image type (fully incongruent), included pairs of arrays where the more numerous array had a smaller cumulative surface area and a smaller convex hull.

The original intention for this study was to create stimuli using the Gebuis and Reynvoet (2011) protocol that exactly matched the numerosities of each trial from the Panamath stimuli. However, because of limitations due to the different ways in which each protocol controls for visual cues, it was not possible to create identical sets of trials. This appears to be because the Gebuis and Reynvoet method struggles to create the intended convex-hull size in trials with small numerosities (the numerosity range attempted was 10-24 as per the Panamath trials). The Gebuis and Reynvoet script contains a warning in the preamble that: "For most designs the program generates stimuli that are not confounded with visual cues. Nevertheless a post hoc analyses to verify whether this is indeed the case is recommended. Especially when small numerosities and large number distances are used, it is unavoidable that strong relations between number and area subtended or circumference arise" (lines 27–32 of script). Post hoc analyses revealed that stimuli created with this script, which were designed to exactly match Panamath numerosities, were indeed confounded with visual cues. As the Gebuis and Reynvoet method generates arrays with different patterns with each run of the script, 20 different attempts were made to create the stimuli, with a post hoc analysis on the visual cues conducted for every attempt. Each time, numerosity was significantly correlated with convex-hull size, with Pearson correlation coefficients ranging from $r = .25$ to $r = .33$. Thus it was not possible to create unconfounded stimuli with the Gebuis and Reynvoet script to match these Panamath stimuli.

Protocol	Num range	Num ratio range	CSA ratio range	CH ratio range
G & R	22–36	0.61–1.64	0.10–11.06	0.45–2.35
Panamath	10–24	0.50–2.00	0.34–1.97	0.56–1.60

Table 3.1: Visual characteristics information for stimuli created with both the Gebuis and Reynvoet (G & R) and Panamath dot comparison protocols, including the range of numerosities represented in the arrays, and the range of the ratios between the two arrays in each trial in terms of numerosity, cumulative surface area (CSA) and convex hull (CH).

Consequently, in order to maximize comparability with existing literature, the Gebuis and Reynvoet (2011) protocol was used as close to its default setting as possible, ensuring that the visual cues were controlled as intended. This involved choosing a slightly larger set of numerosities (22–36 dots) within the typical range from the literature (Dietrich et al., 2015). The task was created with 96 trials, as this has previously been found to be an appropriate number of trials for good reliability (Inglis & Gilmore, 2014). Finally, the yellow dot arrays were always presented on the left of the screen, and the blue dot arrays were presented on the right hand side. The colours were chosen to match the colours of the stimuli created with the Panamath protocol, however did not alternate between the left and the right arrays to match Panamath because the Panamath stimuli had an uneven number of trials of each colour per side. Summaries of the visual characteristics of the arrays created by each protocol are described in Table 3.1. Both of the final stimuli sets were created as close to the default settings of each generation method as possible, and are therefore representative of the standard use of these protocols in the literature.

3.3 Results

The sections below first report an analysis of the characteristics of the dot stimuli produced by each of the protocols. Next, the relationship between participants' performance on each of the protocol conditions, and the test-retest reliability of the trials is explored using Pearson's correlations. Finally, paired t-tests are used to investigate image congruency effects and how they are influenced by divergent visual cue methods used in the literature. Accuracy scores on the dot comparison task were taken as the dependent measure

throughout because accuracy has been shown to be a more reliable measure of performance than w scores or numerical ratio effects (Inglis & Gilmore, 2014).

Ten participants were excluded from the analysis because they did not perform significantly above chance on one or more blocks of the dot comparison task. These participants were excluded because it is not possible to disentangle whether their responses were made in accordance with the aims of the task, by making judgements based on numerosity, or whether they were entirely attending to the visual cues of the stimuli. This is a common exclusion criteria used by other researchers in the field (Inglis et al., 2011; Inglis & Gilmore, 2014; Nys & Content, 2012). This left 47 participants in the final analysis.

3.3.1 Analysis of stimuli

For each of the stimuli, the convex hull and cumulative surface area of the blue and yellow dot arrays was calculated. To obtain the convex hull of each array, the Graham Scan algorithm was used (Graham, 1972).[6] The cumulative surface area of the arrays was calculated by summing the number of coloured pixels in the display. These calculations provided concrete measurements of each trial's visual characteristics using the same method for each protocol.

Analysis of these measurements confirmed that the Panamath protocol created stimuli that did not contain systematic controls for convex-hull size, and therefore convex hull was predictive of numerosity on 37 of the 60 trials. This is represented in Figure 3.2A by the larger number of trials in the upper right and lower left quadrants of the graph, indicating there were significantly more convex-hull congruent trials than convex-hull incongruent trials within the Panamath protocol trials. Consequently, if participants were to complete the task based on convex-hull size judgements alone (with no numerosity processing), they would score 61.67% accuracy, which would

[6]The Graham scan algorithm works by calculating the smallest convex polygon enclosing all the points in the array. The first step in this algorithm is to choose a point O that is interior to the array and to use this as the origin. Next, the input points from the surrounding dots to point O must be sorted in angular order around O. Following this order, a polygon can be formed by joining the points together, eliminating all reflex vertices during the course. The resulting polygon is the convex hull of the array.

result in significantly above chance performance. In contrast, for trials created with the Gebuis and Reynvoet protocol, convex-hull size was predictive of numerosity on exactly half of the trials (48 of 96), as shown in Figure 3.2B by the equal numbers of convex-hull congruent and incongruent trials in each quadrant of the graph. Participants would not be able to perform above chance on these trials using a strategy purely based on convex-hull size. Cumulative surface area was controlled appropriately and was predictive of numerosity on exactly half of the trials for the Gebuis and Reynvoet protocol, and 31 out of 60 trials for the Panamath protocol. The number of cumulative surface area congruent and incongruent trials fell approximately evenly into the diagonally opposing quadrants of the graphs (representing congruent / incongruent boundaries) shown in Figure 3.2C and Figure 3.2D for both protocols.

3.3.2 Relationship between performance across the two protocols

A Pearson correlation showed that individuals' performance on the Gebuis and Reynvoet protocol trials was not significantly correlated with performance on the Panamath protocol trials, $r = .260$, $p = .078$. Although this correlation approached significance, the extremely small R^2 value (.07) demonstrates that only minimal variance in participant's accuracy on Gebuis and Reynvoet protocol trials can be explained by their variation in Panamath scores. This finding indicates that different processes may underlie performance on dot comparison tasks created with different visual controls.

3.3.3 Test-retest reliability

All trials were presented twice within the same testing period, separated by a different block of trials and a short break. A Pearson correlation showed that performance on the first block of trials created using the Gebuis and Reynvoet protocol was significantly correlated with performance on the second, repeated block of these trials, $r = .569$, $p < .001$. In comparison, there was a lower correlation between performances on the first and second blocks of trials created using the Panamath protocol, $r = .286$, $p = .051$.

There were, however, substantially more trials created with the Gebuis and Reynvoet protocol (96 in each block), than trials created with the Pana-

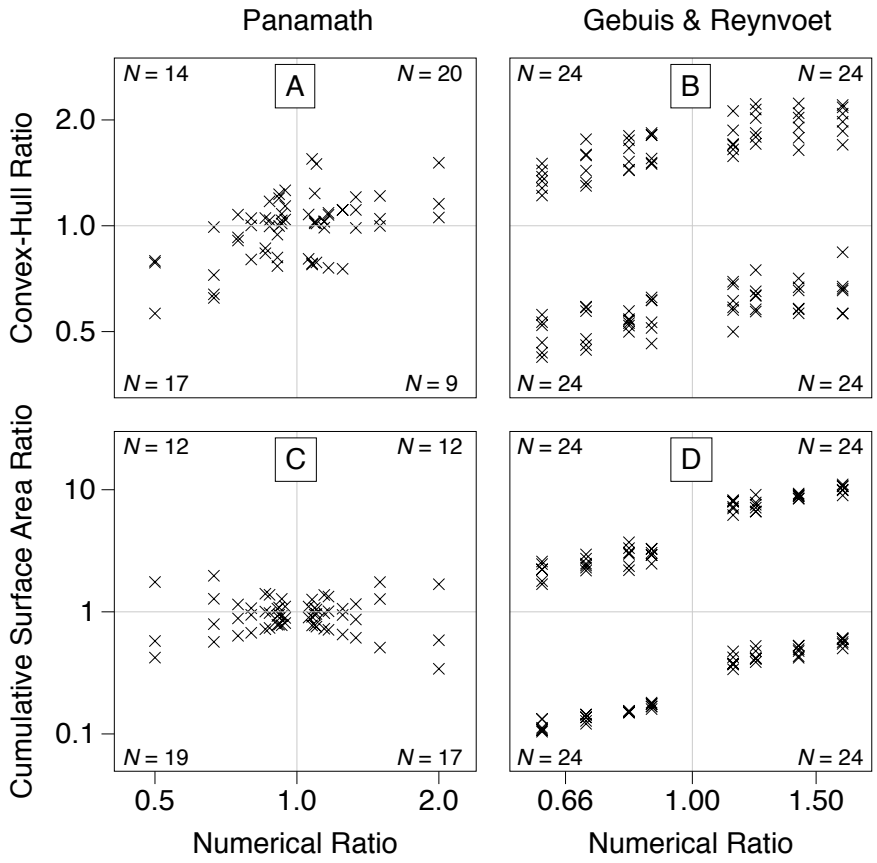


Figure 3.2: Dot comparison trials plotted in terms of the relationships between numerosity ratio and visual cue ratio for each protocol. (a) Numerosity ratio and convex-hull ratio for Panamath trials, (b) numerosity ratio and convex-hull ratio for Gebuis and Reynvoet trials, (c) numerosity ratio and cumulative surface area ratio for Panamath trials, and (d) numerosity ratio and cumulative surface area ratio for Gebuis and Reynvoet trials. The lines that divide the quadrants in this figure define the boundary of congruency effects. For each graph, the upper right and lower left quadrants include congruent trials; the upper left and lower right quadrants include incongruent trials. Axes show a logarithmic scale.

math protocol (60 in each block). To allow for comparability of reliabilities across blocks of trials created with these two different methods, the test-retest reliability of a random subset of 60 Gebuis and Reynvoet protocol trials was also calculated. This analysis was repeated 20 times, each with a different random subset of 60 trials. Pearson correlations showed that the test-retest reliabilities of 60 randomly selected Gebuis and Reynvoet trials were lower than with the full set of 96 trials (Pearson correlation coefficients ranged between .351 and .602, mean $r = .497$, $SD = 0.07$), though these scores nevertheless remained substantially higher than the Panamath test-retest reliability ($r = .286$).

3.3.4 Congruency effects

Using the convex-hull size information obtained with the Graham Scan algorithm (Graham, 1972), and the number of coloured pixels in each array, congruency effects were explored with a 2 (convex-hull size: congruent, incongruent) × 2 (cumulative surface area size: congruent, incongruent) × 2 (protocol: Gebuis & Reynvoet, Panamath) between subjects, by-items ANOVA[7], with mean accuracy per trial as the dependent variable. This resulted in a significant main effect of convex-hull congruency, $F(1, 304) = 317.18$, $p < .001$; participants were more accurate when performing convex-hull congruent trials ($M = 0.88$, $SD = 0.12$), than convex-hull incongruent trials ($M = 0.54$, $SD = 0.18$). There were no significant main effects of cumulative surface area and protocol (see Table 3.2 for descriptive statistics).

Interestingly, the ANOVA resulted in a statistically significant three-way interaction between convex hull, cumulative surface area and protocol, $F(1, 304) = 9.64$, $p = .002$. This interaction was explored further with trials from each protocol separately. For the Gebuis and Reynvoet trials, there was a significant interaction between convex-hull congruency and cumulative surface area congruency, $F(1, 188) = 12.92$, $p < .001$ (Figure 3.3) . This interaction was driven by higher performance on convex-hull incongruent trials when cumulative surface area was congruent ($M = 0.61$, $SD = 0.16$), in comparison to convex-hull and cumulative surface area incongruent trials

[7]A by-items rather than a by-subjects analysis was required here due to the confound between cumulative surface area and convex-hull size in the Panamath stimuli. The cell sizes in a by-subjects analysis would have been highly unbalanced.

Protocol	CH cong		CH incong		CSA cong		CSA incong	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
G & R	0.90	0.06	0.55	0.18	0.76	0.19	0.69	0.25
Panamath	0.86	0.17	0.52	0.25	0.68	0.29	0.79	0.22
Overall	0.88	0.12	0.54	0.18	0.73	0.24	0.73	0.24

Table 3.2: Mean accuracy on trials created with either the Gebuis and Reynvoet (G & R) or the Panamath protocol, categorised into congruent and incongruent conditions (in terms of convex hull and cumulative surface area size).

($M = 0.48$, $SD = 0.19$). In contrast, across convex-hull congruent trials, the cumulative surface area of the arrays did not influence accuracy scores (cumulative surface area congruent: $M = 0.90$, $SD = 0.07$; cumulative surface area incongruent: $M = 0.90$, $SD = 0.05$). This interaction shows that, for Gebuis and Reynvoet protocol trials, convex-hull congruency influenced performance to a greater extent than cumulative surface area congruency.

For the Panamath trials, although the main effect of convex-hull congruency mirrored the same effect found in the Gebuis and Reynvoet trials (higher performance on convex-hull congruent in comparison to convex-hull incongruent trials) a reverse effect was found for cumulative surface area congruency. Participants were more accurate on Panamath cumulative surface area incongruent trials ($M = 0.79$, $SD = 0.22$) than congruent trials ($M = 0.67$, $SD = 0.29$), regardless of convex-hull congruency status. There was no significant interaction between convex-hull size and cumulative surface area in these trials (Figure 3.3).

3.4 Discussion

The present study examined in detail how the differences in two methods of controlling the non-numerical visual cues in dot comparison stimuli influenced task accuracy and reliability. An important finding from this study is that dot comparison tasks created with protocols used by different research groups do not appear to be measuring the same construct. Participants' performance on stimuli created with the Gebuis and Reynvoet (2011) protocol only explained 7% of the variance in their performance on Panamath protocol trials, and performance on trials created with the two protocols was

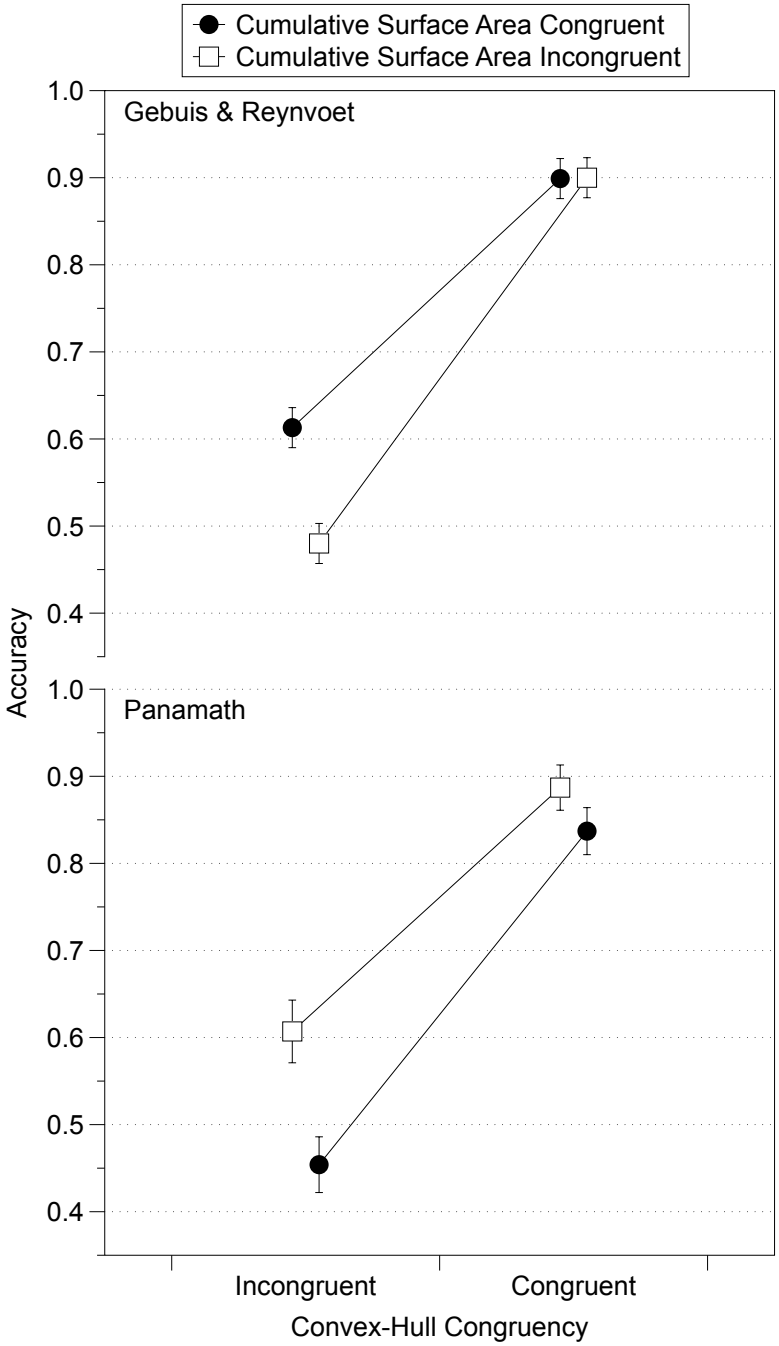


Figure 3.3: Interaction plot of mean accuracy scores calculated in terms of convex hull and cumulative surface area congruency for Gebuis and Reynvoet protocol trials (above) and Panamath protocol trials (below). Error bars represent the standard error of the mean.

not significantly correlated. This finding is in line with a recent study by Smets et al. (2015) which reported a similarly low correlation between performances on stimuli created with the Gebuis and Reynvoet (2011) method, and the Dehaene et al. (2005) script, which follows principles for visual cue controls similar to Panamath. The present result has serious implications for researchers who wish to compare and contrast findings from studies that use different dot comparison task protocols. These tasks appear to be measuring different skills. Although the two sets of trials examined included non-identical numbers of trials and numerosity ranges, if both sets were providing a valid measure of the same underlying construct (i.e. the ANS), one would expect a substantially higher correlation. It must be noted that findings from Panamath protocol trials should be interpreted with caution due to the extremely low immediate test-retest reliability results ($r = .286$). Libertus et al. (2012) similarly found a low test-retest reliability ($r = .22$) for the exactly the same stimuli in their own study, when participants were re-tested with an average of 76.39 days between time one and time two, rather than immediately.

The congruency effects reported here replicate findings from previous research (Barth et al., 2006; Gilmore et al., 2013; Hurewitz et al., 2006; Nys & Content, 2012; Szűcs et al., 2015) and demonstrate that performance on trials created with both the Panamath and Gebuis and Reynvoet protocols is influenced by the congruency status of the visual cues, in particular the convex-hull size. Moreover, the present congruency analysis highlights that measuring and accounting for the convex-hull size as well as cumulative surface area is pivotal to understanding congruency effects. This study shows that participants are significantly more likely to respond correctly to a trial where the larger numerosity has a larger convex hull and larger cumulative surface area, than to a trial where the larger numerosity has a smaller convex hull and smaller cumulative surface area. This result provides clarification on the conflicting findings regarding congruency effects that have been reported in the literature to date; differences are likely due to some researchers failing to consider the convex-hull size of the arrays in their analyses (e.g. Odic, Libertus, et al., 2013; Odic et al., 2014). The present results would not be found if congruency was classified based on total surface area alone. In fact, for trials created with the Panamath protocol, participants performed more accurately on trials where the larger numerosity

had a smaller cumulative surface area. Interestingly, this result is consistent with previous research that has demonstrated that when convex-hull size is kept constant in dot comparison task trials, participants perform better on trials that are incongruent in terms of cumulative surface area, rather than congruent (Gebuis & Reynvoet, 2012a). Given that there is much less range in the convex-hull sizes of the Panamath stimuli, compared to the Gebuis and Reynvoet stimuli, the reverse congruency effect for Panamath trials is in line with this finding. The present results therefore support Gebuis and Reynvoet's (2012a) conclusions that participants do not attend to visual cues independently, but make their judgements by integrating multiple visual cues.

The findings of this study align with recent research demonstrating that methodological differences in tasks believed to measure the ANS have a significant impact on performance (Inglis & Gilmore, 2013; Price et al., 2012; Smets et al., 2014). The findings contribute to the literature by demonstrating that the variation of control for visual cues, a factor many researchers have previously overlooked, has substantial influence on performance patterns. This finding raises issues regarding the underlying cognitive skills that play a role in the completion of dot comparison tasks. Researchers who use dot comparison tasks rarely use identical protocols to previous published studies and consequently work that builds on assumptions from previous literature may be flawed. If researchers are to continue using dot comparison tasks, a standardised protocol must be developed to allow conclusions to be drawn across different studies. Dietrich et al. (2015) have gone some way towards this goal by designing a checklist of methodological aspects to be considered when designing a dot comparison task, however recommendations for the control of many factors remain vague or have not yet been systematically explored.

The implications of the present results also apply to the controversial link between ANS acuity and mathematics ability. As De Smedt et al. (2013) reported, there have been numerous conflicting findings when ANS tasks are presented in a nonsymbolic format using dot arrays. It is difficult to interpret the mixed evidence of existing correlational results when we are still unsure of the processes that contribute to performance on dot comparison tasks. The conflict could be explained, at least in part, by the use of different controls for visual cues.

3.5 Summary of findings

To conclude, this study has demonstrated that there is no correlation between adults' performance on dot comparison trials created by two protocols that use different visual cue controls. Therefore, divergent cognitive processes appear to underlie two non-symbolic comparison tasks that have previously been assumed to measure the acuity of the same construct: the Approximate Number System. The clarification of the existence of visual cue congruency effects supports the hypothesis that the visual characteristics of the stimuli, particularly the convex hull of an array, may inform judgements alongside numerosity information. For incongruent trials, where the visual cues would be an uninformative distractor to the task in hand, individuals may activate inhibitory control mechanisms to account for this and focus on numerosity. Future research should therefore recognise that dot comparison tasks are not pure measures of ANS acuity and should focus on exploring the potential domain general mechanisms that may underlie performance on different versions of this task. Additionally, greater attention should be paid to the reliability of the dot comparison task measures employed as this study has demonstrated that trials created with a widely used protocol have unacceptably low immediate test-retest reliability.

Given the evidence that visual cues substantially influence dot comparison task judgements, and the ANS is hypothesised to extract numerosity information independent from visual information, it is likely that dot comparison tasks do not solely measure ANS acuity, and that other cognitive skills are involved. The next part of this thesis explores the role of inhibition in dot comparison tasks.

Part III

Inhibition in dot comparison tasks

Chapter 4

Inhibition literature review

The previous chapter presented evidence that individuals' judgements of numerosity in dot comparison tasks are substantially influenced by the way in which the visual stimuli are created. The existence of congruency effects demonstrated that visual cues in dot arrays can help or hinder individuals in their relative judgements of quantity. One hypothesis is that for the trials where visual cues are misleading, or incongruent with numerosity, inhibitory control skills are recruited to ignore these cues and regain focus on the demands of the task. This part of the thesis provides an overview of inhibitory control skills, recent research pertaining to the role of inhibition in dot comparison task performance, and finally presents three empirical studies that further explore this link.

4.1 Introduction to inhibition

> "The ability to suppress irrelevant or interfering stimuli or impulses is a fundamental executive function essential for normal thinking processes and ultimately, for successful living" — (Garavan, Ross, & Stein, 1999, p. 8301)

Inhibition, or inhibitory control, is a domain-general executive function skill important for day-to-day life. Inhibition refers to the ability to ignore distracting information and suppress unwanted responses (Dempster, 1992). We often need to suppress irrelevant or distracting stimuli in our daily environment; for example, the sound of a nearby conversation whilst concentrating on work. Inhibitory control skills are important for children

and adults in terms of learning and work productivity, problem solving, and general social skills. Deficiency in inhibitory control processing is related to poorer academic achievement (St Clair-Thompson & Gathercole, 2006), and to clinical disorders such as Attention-Deficit/Hyperactivity Disorder (ADHD), schizophrenia, and Obsessive Compulsive Disorder (OCD) (Friedman & Miyake, 2004).

4.1.1 Subtypes of inhibition

Inhibition can be conceptually distinguished into several categories; in a review of constructs and related measurement paradigms Nigg (2000) specified four types of processing: cognitive inhibition, behavioural inhibition, oculomotor inhibition and interference control inhibition. The first type of inhibition – cognitive inhibition – relates to the ability to suppress intrusive, non-pertinent thoughts, for example, ignoring intrusive thoughts about dinner whilst completing a test. Behavioural inhibition refers to the ability to suppress a prepotent physical response in compliance with changing cues. This is often measured by the 'go/no go' paradigm, whereby participants are required to repeatedly respond to a certain cue, for example by button press, but intermittently withhold this response when a different cue is encountered. A third type of inhibition distinguished by Nigg is oculomotor inhibition, referring to the effortful suppression of reflexive eye-movements, for example, suppressing the urge to look at a novel stimulus that is task irrelevant. Finally, interference control refers to the ability to maintain the execution of a primary motor response in the presence of distracting or competing stimuli pulling for a different response. During interference control, competing stimuli draws attention away from the target response and interferes with the current operations of working memory, in turn slowing the primary motor response (Nigg, 2000).

Interference control is the subtype of inhibition that is most likely to be recruited in a dot comparison task (discussed further in section 4.3), and so all subsequent references to inhibition in this thesis specifically refer to interference control, unless otherwise stated.

4.2 Tasks used to measure interference control inhibition

A classic illustration of a task where inhibitory control skills are recruited is the Stroop task (Stroop, 1935). A standard Stroop task involves stimuli comprising of rows of colour words, e.g. red/ blue/ black, written in different colour inks. Trials can be classified as either congruent or incongruent. Congruent trials involve the colour word matching the ink colour, for example the word 'blue' written in blue ink. Incongruent trials involve the colour word differing from the ink colour, for example the word 'blue' written in red ink. The participant's task is respond to the ink colours without interference from the written word. Consistently it has been found that participants are slower and less accurate when completing incongruent trials in comparison to congruent Stroop task trials (MacLeod, 1991).

Another standard measure of interference control is the Flanker task, developed by Eriksen and Eriksen in 1974. The Flanker task measures individuals' reactions to a target stimulus that is surrounded by a row of either target relevant or target irrelevant stimuli. A common choice of stimuli for this task are arrows (MacLeod, 1991). During an arrows Flanker task, the participant is required to respond to the direction of a central arrow, whilst ignoring the direction that other arrows in the row are pointing.

Multiple variations of these common measures of interference control inhibition have been developed over the years, following the same general principles as the originals, but employing different stimuli, e.g. words, letters, pictures and colours (MacLeod, 1991). Further descriptions of a variety of interference control tasks are provided in Study 4, Chapter 7.

4.3 The role of inhibition in dot comparison tasks

It has been proposed that inhibition ability, specifically interference control skills, play an important role in dot comparison performance as a result of the way dot stimuli are created (Szűcs et al., 2015). As described in Part II of this thesis, in order to ensure that participants solve dot comparison tasks on the basis of the numerosity of the arrays, rather than visual characteristics, such as dot size or convex hull, dot comparison tasks typically consist of both congruent and incongruent trials. On congruent trials, visual cues such as

the average dot size and convex hull of the array are positively correlated with numerosity i.e. the array with more dots is made up of larger dots and covers a greater area. Conversely, on incongruent trials, average dot size and the convex hull of the array are negatively correlated with numerosity i.e. the array with fewer dots is made up of larger dots and covers a greater area. Further in-depth discussion on the topic of visual cue control in non-symbolic stimuli is provided in Chapter 2.

Some researchers have proposed that the congruency categories of dot comparison task trials are comparable to the different congruency categories present in Stroop task trials (Gilmore et al., 2013; Nys & Content, 2012; Szűcs et al., 2015). Gilmore et al. (2013) suggested that for a participant to respond accurately to an incongruent dot comparison task trial they must inhibit the irrelevant and misleading visual information, such as dot size and convex hull, and respond solely based on numerosity estimations. There is a wealth of evidence to show that participants perform significantly slower and less accurately on trials where the continuous visual variables are not predictive of numerosity (Barth et al., 2006; Gebuis, Kadosh, de Haan, & Henik, 2009; Gilmore et al., 2013; Hurewitz et al., 2006; Nys & Content, 2012). Gilmore et al. (2013) proposed that this is likely to be due to the added inhibitory control demand required specifically for incongruent trials, in order to ignore the misleading visual cues.

4.4 Inhibition as a mediator in the relationship between dot comparison performance and mathematics achievement

Due to the above congruency effects, Gilmore and colleagues (2013) proposed that inhibitory control may be pertinent to the debate surrounding the relationship between the ANS and formal mathematics achievement. It is well documented within the psychology and mathematics education literatures that there is a relationship between inhibitory control and formal mathematics ability (e.g. Blair & Razza, 2007; Bull & Scerif, 2001; Espy et al., 2004; St Clair-Thompson & Gathercole, 2001). Individuals with better inhibition skills also tend to perform better on tasks measuring mathematical ability (see Chapter 1, Section 1.1.1). In line with this, Gilmore et al.

(2013) found that children's formal mathematics achievement scores were only correlated with performance on incongruent dot comparison task trials, and not with congruent trials. Gilmore et al. therefore proposed that the correlation often observed between mathematics achievement and ANS acuity may arise from mutual correlations with inhibitory control. Indeed Gilmore et al. (2013) also reported that 7 to 10 year olds' overall dot comparison performance scores no longer significantly predicted mathematics achievement scores once inhibition skills were accounted for. Supporting evidence for this proposal is also demonstrated in Fuhs and McNeil's (2013) study with low social-economic-status (SES) preschoolers. Fuhs and McNeil (2013) found that dot comparison task performance was no longer a borderline predictor of mathematics achievement once inhibition task scores were controlled for. Recently, Cappelletti et al. (2014) have suggested that a decline in dot comparison task performance in an older population may reflect deterioration of inhibitory processes rather than impoverished numerical skills. Older participants were particularly impaired on tasks that required the inhibition of visual information incongruent to numerosity, and moreover this difficulty was correlated with poorer inhibitory control performance on a classic Stroop task.

4.5 A competing processes account

Put together, this provides strong evidence for a competing processes account of dot comparison task performance (Fuhs & McNeil, 2013; Gilmore et al., 2013; Nys & Content, 2012). These authors suggested that both the ANS and other competing processes, driven by visual cues, influence accuracy on non-symbolic comparison tasks. According to this account, when a participant is faced with an individual dot comparison trial, they engage their ANS to judge the difference between the two dot arrays. However, at the same time the visual characteristics of the stimuli such as dot size and convex-hull size may interfere with this process, and in some cases participants may have to inhibit a response based on these visual characteristics in order to respond correctly.

Nevertheless, there has been some resistance to the proposal of the competing processes account. Some researchers do not find congruency effects using their dot comparison stimuli (Odic, Libertus, et al., 2013; Odic et al.,

2014) and therefore oppose the view that inhibition is involved. However, as was established in the study presented in Chapter 3, the lack of congruency effects reported in these studies is likely due to the lack of measurement of multiple visual cues. When convex-hull size is taken in to account, congruency effects are clear (see Chapter 3, Section 3.3.4).

Nonetheless, a recent paper by Keller and Libertus (2015) investigated whether inhibitory control could explain the link between ANS task performance and mathematics abilities in preschoolers from middle- to high- SES backgrounds. Keller and Libertus (2015) found no difference in participants' accuracy scores on different visual cue congruency conditions (although they did not account for convex-hull size), and consequently collapsed performance across all trial types for the analyses. In conflict with previous findings (Fuhs & McNeil, 2013; Gilmore et al., 2013; Cappelletti et al., 2014), Keller and Libertus found no significant correlation between dot comparison task performance and an interference control measure of inhibition (NEPSY-II subtest, a standardised measure of inhibition). The authors did, however, report a significant relationship between individual differences in children's dot comparison task performance and their performance on a standardised measure of mathematics achievement. Interestingly, Keller and Libertus found that this correlation remained significant, albeit reduced, when individual differences in inhibition task performance were controlled for. This finding is in direct conflict with Gilmore et al.'s (2013) study which used exactly the same inhibitory control task to explore this relationship. Importantly, Gilmore et al.'s dot comparison task used the Gebuis and Reynvoet (2011) method of controlling visual cues in their stimuli, whereas Keller and Libertus used a method whereby convex-hull size was uncontrolled. If Keller and Libertus's stimuli happened to contain a confound between convex-hull size and numerosity, as found in the stimuli used by Libertus et al. (2012) (Chapter 3, Section, 3.3.1), then participants may not have needed to inhibit visual cues to perform successfully on the task. It is possible that participants could have performed successfully by responding to the larger visual cues. Therefore, the correlation between dot comparison task performance and mathematics reported by Keller and Libertus (2015) may be due to a mutual correlation with the visuo-spatial processing skills required to weigh up visual cues.

In addition to different stimuli generation methods, Keller and Libertus

(2015) also used a different analysis technique to Gilmore et al. (2013) to examine inhibition ability. The NEPSY-II Inhibition task requires participants to first name rows of intermixed circles and squares correctly ('naming' condition), and then do the same task responding with the alternative shape's name, e.g. responding "square" when it is a circle and "circle" when it is a square ('inhibition' condition). Keller and Libertus assessed inhibition performance using a combined contrast score of the naming and inhibition elements from this task. This measure of performance reduces both the naming and the inhibition elements to standardised scores, and then uses these two standardised scores to form a final contrast score as analysed by Keller and Libertus (2015). In comparison, Gilmore et al. (2013) assessed the influence of both the naming and inhibition elements of this task separately to provide a more sensitive measure that accounts for overall levels of performance. Therefore, it is possible that the divergent results obtained by Keller and Libertus (2015) and Gilmore et al. (2013) are due to methodological differences in dot comparison task procedures and inhibition task analysis techniques. More research is required to investigate the role of inhibition in dot comparison tasks further.

4.6 Summary

The findings from Part II of this thesis suggest that the visual characteristics of dot comparison tasks substantially influence individual's judgements about numerosity. This provides initial insight into the processing that dot comparison tasks involve, but does not explain the cognitive mechanisms with which individuals use this visual information to make their quantity judgements. Here, the evidence reviewed suggests that inhibitory control skills may play a role in dot comparison task judgements in order to compensate for visual cues that are incongruent with the non-symbolic quantity represented. A competing processes account is described which proposes that both the ANS and inhibitory control skills contribute to dot comparison task performance scores. Although several studies have found that inhibition plays a significant role in dot comparison task performance, and even mediates the relationship between ANS acuity and mathematics achievement, contradictory evidence also exists. Thus, further research is needed to examine the role of inhibition in non-symbolic comparison.

Chapter 5

Set size study (Study 2)

The following study investigated whether inhibition is likely to be recruited during dot comparison task judgements through an analysis of the effects of changing visual cue salience. If more obvious, or salient, visual cues cause bigger differences between accuracy on congruent trials in comparison to incongruent trials, it follows that this could be due to difficulties inhibiting misleading visual cues on incongruent trials. This hypothesis was explored through an investigation of how changes to the magnitudes of numerosities affects the relative salience of visual characteristics, and in turn influences congruency effects. The results are discussed with reference to the competing processes inhibition-based account of performance.[1]

5.1 Introduction

Section 1.3.2 of the literature review in Chapter 1 of this thesis introduced the issue of methodological irregularities within published ANS task studies. One of the most under-investigated aspects of dot comparison task methodology relates to the range of numerosities, or the set size of the arrays, represented in dot comparison task trials. There is no consensus as to the appropriate range of numerosities that should be included in a task, and dot comparison stimuli can represent as few as four dots (Libertus et al., 2011), to as many as 70 (Inglis et al., 2011). A review by De Smedt et al. (2013)

[1]The data presented in this chapter are published in *ZDM Mathematics Education* (Clayton & Gilmore, 2014).

highlighted this lack of standardisation and suggested that inconsistencies in set size could play a role in the explanation of contrasting results within the dot comparison literature.

The standard model of the ANS is based on the assumption that variations in the absolute magnitude of dot arrays should not affect ANS judgements, so long as the ratio between the two arrays remains constant. The model predicts that performance is only influenced by the numerosity ratios of the trials and the individual's ANS acuity (Barth et al., 2005; Dehaene, 1997). Therefore, according to the standard model of ANS processing, participants should perform equally on trials that have the same ratio between the to-be-compared numerosities, irrespective of the magnitude of the arrays, e.g. performance on a 7 vs. 10 trial is predicted to be equal to performance on a 70 vs. 100 trial.

Acceptance of this assumption may have led to the limited amount of research into the effects of varied set sizes in dot comparison tasks. The researchers who have reported the influence of variation in set size on performance have only done so through the analysis of existing data sets, and not through a planned systematic study of set size. As described in previously in Section 1.3.2, Barth et al. (2008) compared different individuals' performances on two dot comparison tasks with marginally different set sizes (16–56 vs. 5–40 dots) and found no differences in accuracy scores across the two studies. In contrast, a study by Revkin and colleagues reported higher performance for small sets (1–4) in comparison to larger sets (10–40) (Revkin et al., 2008). However, the authors suggested that the process of subitizing very small sets of items is characterised by different underlying mechanisms to ANS representations used for larger sets of items (Revkin et al., 2008), and so this result does not inform the question of set size effects within ANS tasks. Therefore the effect of set size on dot comparison task performance remains unknown.

As discussed in the review of inhibition literature provided in Chapter 4, previous research has suggested that inhibitory control skills may play an important role in dot comparison performance (Cappelletti et al., 2014; Fuhs & McNeil, 2013; Gilmore et al., 2013; Nys & Content, 2012; Szűcs et al., 2015). According to the competing processes account, participants may use both ANS processing and inhibitory control to complete a dot comparison task (Fuhs & McNeil, 2013; Gilmore et al., 2013; Nys & Content, 2012).

Although recent studies have indicated the involvement of inhibitory control skills in general, questions still remain regarding the characteristics of dot comparison trials that may increase inhibitory control demands. In order to understand the ways in which inhibition is involved in solving incongruent dot comparison task trials, it is necessary to further understand how changes to visual cues influence task performance. This is particularly important given that different researchers have employed divergent methods to create stimuli (Dietrich et al., 2015).

The aim of the present study was to investigate the characteristics that have the potential to influence congruency effects, and explore whether particular visual cues are more difficult to inhibit than others. This study explored three factors: the numerical ratio between the arrays, the absolute set size of the quantities represented, and the visual characteristics of the arrays, specifically average dot size[2] and convex hull. If both ANS acuity and inhibitory control skills influence performance on dot comparison tasks, as proposed by the competing processes account, then participants' accuracy scores are hypothesised to be related to all three factors. Numerical ratio would influence performance due to the approximate nature of ANS representations, according to Dehaene (1997). Set size and consequential changes in visual characteristics of the stimuli would influence performance by varying the inhibitory control demands of the incongruent trials. It was hypothesised that the salience of the visual cues in an array would increase with absolute set size. In particular, convex-hull size is likely to become more salient with increasing numerosity as the density of the dots increases and creates a more prominent perimeter (see Figure 5.1 for an example).

[2]Note that the influence of average dot size was analysed in this study, and all following studies reported in this thesis. This is in contrast to Study 1, which assessed the influence of cumulative surface area. As mentioned in Chapter 2, Section 2.2, there is no substantial benefit in investigating the effects of more than one of these 'dot size' cues, as they are highly correlated. Study 1 differed from the present and remaining studies in investigating cumulative surface area because a main aim of Study 1 was to draw comparisons between congruency effects measured with Panamath stimuli, and with Gebuis and Reynvoet stimuli. Studies that use Panamath stimuli more often report congruency status in terms of the relationship between numerosity and cumulative surface area, rather than dot size (e.g. Libertus et al., 2012), thus the same factor was explored in Study 1 to aid comparisons with studies that have used the Panamath protocol. The remaining studies reported in this thesis use the Gebuis and Reynvoet protocol to create stimuli, and analyses assess the influence of average dot size, rather than cumulative surface area, because this is a variable commonly reported in papers using protocols other than Panamath.

For an incongruent trial, this is hypothesised to place a greater demand on inhibitory control skills to disregard the more noticeable interfering information (convex-hull size) and focus on numerosity.

A number of predictions were made regarding participants' performance on in this study. Firstly, overall accuracy was predicted to decrease as the set size of the arrays increased. Second, convex-hull incongruent trials were predicted be more challenging than convex-hull congruent trials. Third, an interaction between this congruency effect and set size was predicted. Participants were hypothesised to perform less accurately on incongruent trials that were made up of larger absolute numerosities in comparison to smaller numerosities, whilst numerosity ratios remained constant. Finally, trials that were incongruent in terms of dot size were predicted to be more challenging than dot-size congruent trials, however no predictions were made regarding interaction effects, as it is unclear whether dot size congruency effects would vary with increasing set size.

5.2 Method

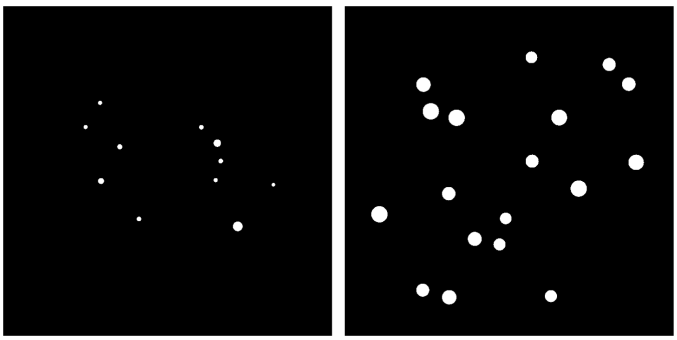
5.2.1 Participants

Participants were 44 children (22 male) aged between 7 and 9 years (M = 8.3, SD = 0.59 years). Children were tested in a quiet area of their school and were given a certificate for taking part. This study was approved by the Loughborough University Ethics Approvals (Human Participants) Sub-Committee.

5.2.2 Task

Participants completed a dot comparison task. Stimuli were arrays of white dots presented on a black background. The dots were generated using Gebuis and Reynvoet's (2011) method, which creates each pair of dot arrays four times with different visual characteristics in terms of average dot size and convex hull, resulting in four image types (fully congruent; dot-size congruent, convex-hull incongruent; dot-size incongruent, convex-hull congruent; fully incongruent, described in Section 2.3, and illustrated in Figure 2.4, Chapter 2). There were four set size conditions: small, medium, large and very large. The small numerosities ranged from 10 to 19, the larger sets were

(a)



(b)

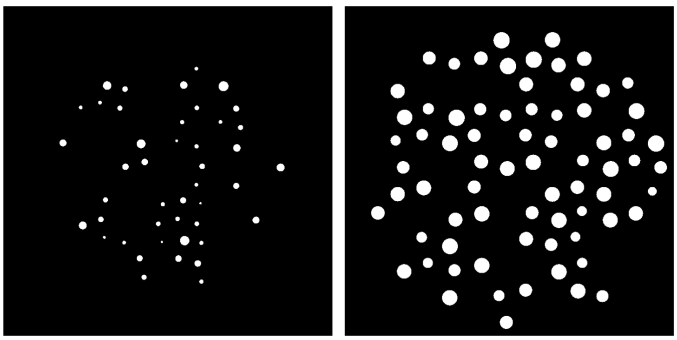


Figure 5.1: Examples of the same 0.61 numerosity ratio trial in a) the smallest set size (11 vs. 18 dots) and b) the largest set size (44 vs. 72 dots). The convex-hull size appears more prominent in the larger set size than the smaller set size arrays.

2, 3, and 4 times as large respectively. The ratios between the numerosities displayed in the arrays were 0.53, 0.61, 0.71, 0.81, 0.88 and 0.93. The stimuli were presented simultaneously, side-by-side on a 15" laptop display. Each trial began with a fixation point (600 ms) followed by the presentation of the two arrays (600 ms) and finally a screen with a question mark, which was displayed until the participant responded. Participants were asked to indicate which array was more numerous using left and right keys marked on the keyboard. There were eight practice trials and 184 experimental trials. The task lasted approximately 5 minutes.

5.3 Analysis

Seven children were excluded because their dot comparison task performance was not significantly above chance. This left 37 participants in the analysis.

For the purposes of clarity, details of each part of the data analysis are provided alongside the corresponding results in the following section.

5.4 Results

First the effects of set size and the stimuli congruency status on performance were explored. Participants' accuracy data were subjected to a 4 (set size: small, medium, large, very large) × 2 (convex hull: congruent, incongruent) × 2 (average dot size: congruent, incongruent) within-subjects Analysis of Variance (ANOVA). As predicted, accuracy scores were highest in the small set size condition, M= 64.0%, and declined with increasing numerosity, Ms= 62.2%, 61.8% and 60.2% for the medium, large and very large conditions respectively. This represented a significant linear trend, $F(1, 36) = 6.6$, $p = .014$, $\eta_p^2 = .16$.

Accuracy was significantly higher for trials that were congruent in terms of convex hull ($M = 80.8\%$) compared to incongruent trials ($M = 43.3\%$), $F(1, 36) = 158.18$, $p < .001$, $\eta_p^2 = .82$. Similarly, accuracy was significantly higher for trials that were congruent in terms of dot size ($M = 69.8\%$) than incongruent trials ($M = 54.2\%$), $F(1, 36) = 14.43$, $p < .001$, $\eta_p^2 = .29$. Set size significantly interacted with both convex-hull congruency, $F(3, 108) = 37.18$, $p < .001$, $\eta_p^2 = .51$ and dot-size congruency, $F(3, 108) = 5.92$, $p < .001$, $\eta_p^2 = .14$. As set size increased, the effect of convex-hull congruency

increased, shown in Figure 5.2. In comparison, as set size increased, the effect of dot-size congruency decreased, shown in Figure 5.2. Notably, performance drops significantly below chance on the convex-hull incongruent trials of medium, $t(36) = -3.37$, $p = .002$, large, $t(36) = -3.67$, $p = .001$, and very large set sizes, $t(36) = -5.36$, $p < .001$.

In addition to this analysis, a series of binary logistic regressions were conducted to investigate more sensitively how the ratio between cues on each trial (numerosity, convex hull, dot size) affected individuals' performance, and how this differed between smaller and larger set size trials. The average dot size of each array was calculated as the total number of white pixels divided by the number of dots in each image. The convex-hull size was calculated using the Graham Scan Algorithm (Graham, 1972), described further in Chapter 3, Section 3.3.1. Using these values, for each trial, the ratios between the two images in terms of average dot size, convex-hull size and numerosity were calculated. Then, for each individual participant, a binary logistic regression was conducted predicting trial response using the ratios between the trial's two numerosities, the two convex hulls, and the two mean dot sizes. This yielded odds ratios for convex-hull size, dot size and numerosity. These odds ratios were calculated for the full set of trials, and then calculated separately for trials that included smaller set sizes (small and medium set size groups) and larger set sizes (large and very large groups). A Wilcoxon Signed-ranks test was used to compare the odds ratios derived from these different set size groups.

An odds ratio significantly greater than 1 indicates that the given predictor has had an effect on the individual's comparison performance. Overall, the odds ratios for numerosity ratio were higher than 1 for every participant, suggesting that all the participants used numerosity information to some extent to complete the task. A Wilcoxon signed-ranks test showed no significant difference between odds ratios for the numerosity ratio for smaller ($Mdn = 2.99$) and larger ($Mdn = 3.33$) set sizes, $Z = .309$, $p = .757$, suggesting that participants focused on numerosity irrespective of the set size.

The odds ratios for the convex-hull ratio did, however, vary by set size. A Wilcoxon Signed-ranks test showed that odds ratios for convex-hull ratio were significantly lower, $Mdn = 2.69$, for smaller compared to larger, $Mdn = 9.40$, set sizes, $Z = 5.21$, $p < .001$. For dot-size ratio, there was no difference

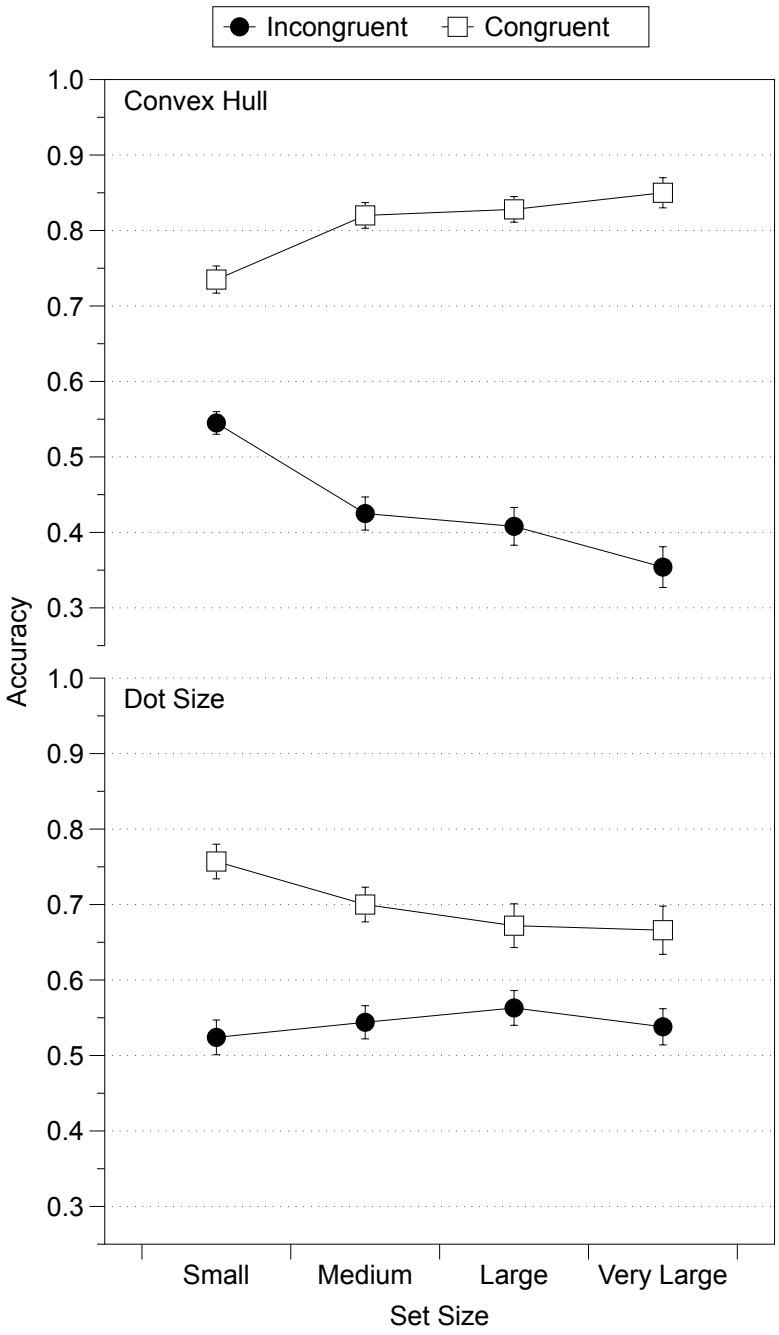


Figure 5.2: Above: Accuracy scores for convex-hull congruent and incongruent trials with small, medium, large and very large set sizes. Below: Accuracy scores for dot-size congruent and incongruent trials with small, medium, large and very large set sizes. Error bars represent the standard error of the mean.

between odds ratios for smaller, $Mdn = 1.17$, and larger, $Mdn = 1.20$, set size trials, $Z = 1.74$, $p = .083$. This pattern of odds ratios is consistent with the picture that emerged from the ANOVA analysis. The findings suggest that the impact of visual characteristics in dot arrays varies with set size. Participants are more influenced by the convex-hull size of the array on trials with larger set sizes in comparison to trials with smaller set sizes of identical numerical ratios.

5.5 Discussion

The aim of this study was to investigate how dot comparison task performance was influenced by individual differences in ANS acuities but also by more wide-ranging domain-general cognitive skills. Specifically, this study examined whether dot comparison accuracy scores reflected participants' ANS acuity alone, indicated by significant effects of numerosity ratio, or whether the visual characteristics of the arrays also impacted on performance, suggesting the involvement of additional processing. To do this, the ratios between the numerosities, and the absolute set sizes of the to-be-compared dot arrays were manipulated, which in turn affected the visual characteristics of the arrays.

As predicted, the present study showed that as the numerosities represented in the stimuli increased, overall accuracy scores decreased. Furthermore, the visual cues that participants attended to most varied by set size. Specifically, as set size increased, participants were more influenced by the convex hull of the arrays, and less so by the average dot size. To elaborate, participants performed more accurately on trials where the convex-hull size was predictive of numerosity (i.e. convex-hull congruent trials), and less accurately when it was incongruent with numerosity (i.e. convex-hull incongruent trials). This was the case across all set size conditions, though the difference was greater in larger set sizes in comparison to smaller set sizes. In fact, performance was significantly below chance on the medium, large and very large set size convex-hull incongruent trials, suggesting that participants found it particularly difficult to ignore convex-hull cues on these trials. Similarly, across all set size conditions, participants performed more accurately on trials where the average dot size was predictive of numerosity (i.e. dot-size congruent trials), and less accurately when it was incongru-

ent with numerosity (i.e. dot-size incongruent trials). In contrast to the convex-hull set size effect, this difference was greater in smaller set sizes in comparison to larger set sizes. This result shows that different visual cues had more impact on performance depending on the absolute set size of the dot array. Importantly, regression analyses using more precise measures of the visual cues confirmed these findings and additionally demonstrated that accuracy scores were influenced not only by visual cue processing, but also by numerosity processing.

The present results have implications for ANS theory. First, the standard model of the ANS (Dehaene, 1997) struggles to account for these findings. This model claims that individuals' ANS precision and the ratio between the numerosities in each trial should be the only two predictors of accuracy in dot comparison tasks (Barth et al., 2005; Dehaene, 1997). Although results show that all participants focus on numerosity to some extent to complete the task, they also show that set size and visual cues, such as dot size and convex-hull size, interfere with task performance. This reveals that task success depends on more than just ANS processing, and other cognitive skills are recruited to process the visual cues in the stimuli.

The present results are in line with the results reported in Study 1, Chapter 3, of this thesis, and support the view of Gebuis and colleagues who argue that the visual characteristics of dot comparison tasks are of pivotal importance to performance on dot comparison tasks (Gebuis & Gevers, 2011; Gebuis & Reynvoet, 2012a, 2012b). Interestingly, results demonstrated that the congruency status of convex-hull size had a stronger influence on individuals' performance than average dot size, as reflected by the larger overall congruency effects reported in the ANOVA. Correspondingly, as set size varied, the influence of convex hull on participants' accuracy scores varied significantly. However, despite the significant ANOVA interaction, the odds ratio analyses showed the influence of average dot size to be less prominent. This finding corresponds with the results from Study 1 that showed weaker congruency effects caused by cumulative surface area (a cue highly correlated with average dot size), in comparison to convex hull. The present findings support Gebuis and colleagues' proposal that numerosity judgements can be made as a function of weighing up multiple visual cues simultaneously (Gebuis & Gevers, 2011; Gebuis & Reynvoet, 2012a, 2012b). The results of the present odds ratio analysis, however, suggest that ANS representations

may in fact be employed in numerosity judgements *alongside* visual cue processing. However, because this study was unable to distinguish the relative important of each of these factors, this is a hypothesis that warrants further investigation.

The findings from this study provide support for the competing processes account of dot comparison task performance (Fuhs & McNeil, 2013; Gilmore et al., 2013; Nys & Content, 2012). Gilmore et al. (2013) suggested that the successful completion of a dot comparison task relies partially on ANS abilities and partially on the ability to inhibit salient visual features of the array. The present results are consistent with the suggestion that participants do use their ANS to perform dot comparison tasks, and show that the congruency of the visual stimuli, in terms of dot size and convex-hull size, also influences performance. Participants found trials where it was necessary to inhibit the incongruent visual characteristics of an array significantly more difficult than congruent trials. This is not to suggest that ANS processing itself is influenced by interfering processes, but rather that the interference competes with ANS processing, and consequently influences dot comparison performance.

From this evidence, it seems likely that dot comparison tasks may also measure individual differences in interference control. Under this view, participants with better inhibitory control skills are likely to perform more accurately on dot comparison task trials that contain incongruent visual cue information, in comparison to participants with poorer inhibitory control skills. This finding is consistent with previous research that has found a correlation between performance on dot comparison trials where visual cue information was incongruent with numerosity, and performance on inhibitory control tasks (Cappelletti et al., 2014; Fuhs & McNeil, 2013). Following this, further research should focus on the extent to which performance on dot comparison tasks can be accounted for by inhibitory control skills rather than individual differences in ANS acuity.

The findings from the present study also have significant methodological implications for dot comparison task research and underscore the significance of the many procedural differences within the literature. Currently, there is no consensus concerning the range of numerosities included in comparison tasks and many studies that are cited and reviewed in the literature involve diverse ranges of set sizes. Similarly, as highlighted in Chapter 2, there

are no established recommendations on how visual characteristics should be controlled in the stimuli. Many researchers only control for the cumulative surface area and average size of the dots (e.g. Halberda et al., 2008; Halberda & Feigenson, 2008; Halberda et al., 2012; Hellgren et al., 2013; Libertus et al., 2011, 2012, 2013a, 2013b; Mazzocco et al., 2011a, 2011b; Odic et al., 2014; Odic, Libertus, et al., 2013) and so fail to investigate how performance is influenced by important visual cues such as convex hull. As shown by the present results, variations in set size, convex hull and average dot size all influence accuracy scores and so should be considered carefully when designing, analysing and comparing non-symbolic dot comparison experiments.

5.6 Summary of findings

The results of this study demonstrated that children's accuracy scores on a dot comparison task designed to measure the ANS were influenced not only by individual differences in ANS acuity, but also by the size of the numerosities involved and the visual characteristics of the stimuli. Even with numerosity ratios held constant, performance was found to decline as the set size of the stimuli increased. To illustrate, a 70 vs. 100 dot trial may be more difficult than a 7 vs. 10 dot trial, in conflict with predictions from the dominant model of the ANS. Results follow a pattern consistent with the hypothesis that inhibitory control may have been recruited to account for visual cues that were incongruent with numerosity information. As set size increased, visual cues necessarily altered, and congruency effects suggested that convex hull became harder to inhibit. This finding strengthens evidence for the crucial role of inhibition in dot comparison tasks, although more research is needed to support this hypothesis.

Chapter 6

Frequency of conflict task (Study 3)

Studies 1 and 2 presented in Chapters 3 and 5 of this thesis have provided evidence of the important role that visual characteristics play in forming dot comparison task judgements. The results from Study 2 showed that both numerosity and visual cue judgements influenced individuals' performance, and therefore it is possible that both ANS processing and inhibition skills are involved in the completion of dot comparison tasks. Study 3, presented here, builds on these findings to provide additional support for the view that cognitive skills other than ANS acuity influence dot comparison task accuracy.

6.1 Introduction

The competing processes hypothesis (introduced in Section 4.5, Chapter 4) proposes that visual cues in dot array stimuli, such as the average dot size and convex-hull size, may interfere with ANS processing of numerosity in dot comparison tasks. Consequently, for trials where the numerosity of the array is incongruent with the size of the visual characteristics, inhibition skills may be recruited to override the misleading visual cue interference and to focus on numerosity. Although this theory follows from the existence of congruency effects, there is mixed support for the role of inhibition in dot comparison tasks (Cappelletti et al., 2014; Fuhs & McNeil, 2013; Gilmore et al., 2013; Nys & Content, 2012, but see Keller & Libertus, 2015, for an

alternative view).

Dominant theories of ANS processing, such as the Dehaene (1997) model, suggest that an individual's performance on a dot comparison task is only influenced by the ratio difference between the quantities represented and the individual's ANS acuity. Therefore, this model would predict that dot comparison task trials are processed on an individual trial-by-trial basis, without interference from previous trials. In contrast, research from the inhibition literature has found that participants show less interference on incongruent inhibition task trials if these are frequent relative to congruent trials (Botvinick, Braver, Barch, Carter, & Cohen, 2001; Lindsay & Jacoby, 1994; Logan, Zbrodoff, & Williamson, 1984; Tzelgov, Henik, & Berger, 1992). For example, Tzelgov et al. (1992) demonstrated that the response time difference between incongruent and neutral Stroop task trials decreased as the proportion of incongruent trials per block increased. Furthermore, Henik, Bibi, Yanai, and Tzelgov (1997) found that participants showed more interference on the initial one or two trials in an incongruent Stroop task than on subsequent trials. Following these findings, if inhibition is substantially involved in dot comparison task processing, one would expect performance to vary depending on the congruency status of the preceding trials.

Braver's (2012) Dual Mechanisms of Control (DMC) framework provides an explanation for this interesting pattern of results within the inhibition domain. Braver (2012) suggested that cognitive control operates via two distinct operating modes: 'proactive control' and 'reactive control'. Proactive control can be seen as the 'early selection' of a response, keeping goal relevant information active in mind throughout. Reactive control can be seen as 'late correction', with responses operating 'just in time' after high interference is detected. Braver suggests that when expectancy levels are high, proactive instead of reactive control is recruited. Therefore, in an inhibition task with a high proportion of incongruent trials relative to congruent trials, proactive control may lead to faster processing as the goal relevant information is activated and maintained throughout the task.

The design of the present study was guided by this insight from the inhibition literature. In order to provide evidence as to whether inhibitory control is recruited when completing incongruent dot comparison task trials, the frequency of conflict between congruent and incongruent dot comparison task trials was manipulated. The aim of this study was to establish whether

dot comparison task performance is influenced by the processing of previous trials in the same way as inhibition task performance. Given the previous evidence of this pattern of performance in classic inhibition task settings (Botvinick et al., 2001; Lindsay & Jacoby, 1994; Logan et al., 1984; Tzelgov et al., 1992), if performance on incongruent dot comparison trials improved when preceded by multiple similarly incongruent trials, in comparison to when preceded by a block of congruent trials, this would provide compelling evidence for the role of inhibition.

6.2 Method

6.2.1 Participants

Participants were 12 adults (4 male) aged between 20 and 38 years (M = 24.73 SD = 6.05) from Loughborough University. Participants were given an inconvenience allowance of £3 to take part and were tested individually in a quiet room. This study was approved by the Loughborough University Ethics Approvals (Human Participants) Sub-Committee.

6.2.2 Task

Participants completed a dot comparison task on a computer, during which they were required to select the more numerous of two dot arrays. The two arrays consisted of white dots on a black background and were presented simultaneously, side-by-side on a 15" laptop screen. Participants were asked to select which array was more numerous using left and right keys marked on the keyboard. Each trial began with a fixation point (600 ms), followed by the presentation of the two arrays (600 ms) and finally a black screen with a white '?' in the centre was presented until a response was given.

The ratios between the numerosities of the arrays were 0.7, 0.8, 0.9, and 0.95. Numerosities in the set size ranged from 53 to 76. The dots were created using the Gebuis and Reynvoet (2011) method to control for continuous quantity variables. This created four image types (see Section 2.3 and Figure 2.4 for further details), but only the fully congruent and fully incongruent image types (images 1 and 4 in Figure 2.4) were used in this study. The fully congruent trials included pairs of arrays where the more numerous array contained larger dots and had a larger convex hull. The

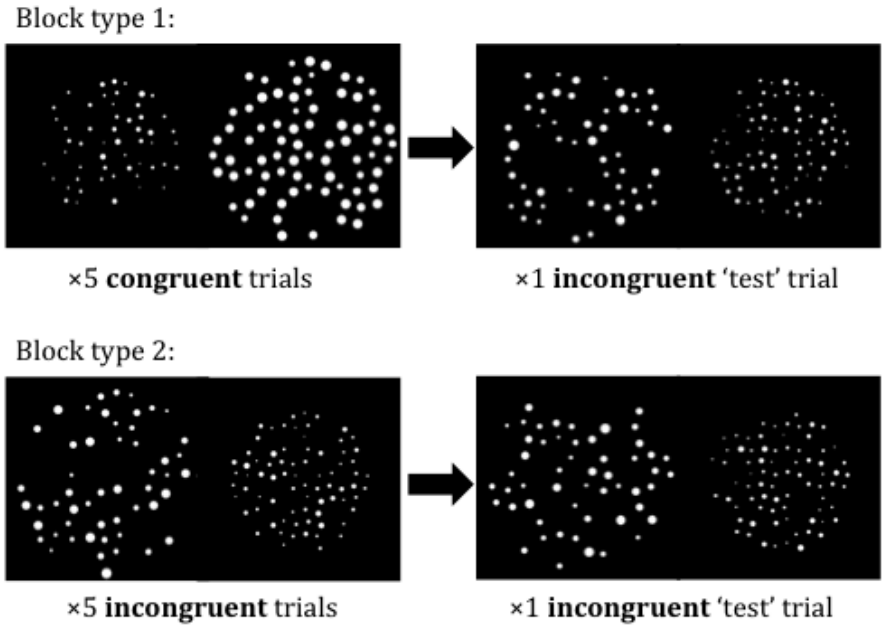


Figure 6.1: The order of dot comparison trials in terms of congruency status for block type 1 and block type 2.

fully incongruent trials included pairs of arrays where the more numerous array contained smaller dots and had a smaller convex hull.

Trial order was manipulated in terms of the frequency of conflict between the congruent and incongruent trials. This was done in two different patterns of trial order that I will refer to as block type 1 and block type 2. In block type 1, participants completed a set of five sequential congruent trials, followed by one incongruent 'test' trial. In block type 2, participants completed a set of five sequential incongruent trials followed by one incongruent 'test' trial. Figure 6.1 presents a visual representation of trial order for each block type. Participants completed 40 blocks of each block type, presented in a random order, totalling 480 trials. Of these trials, 80 were 'test' trials used in the analysis. Participants were additionally given eight practice trials to begin. The task took approximately 20 minutes to complete.

6.3 Analysis

The dependent variable for this study was mean accuracy scores on the incongruent test trials. Participants' mean accuracy scores on these trials were subjected to a 2 (block type: block type 1, block type 2) $\times$ 4 (ratio: 0.7, 0.8, 0.9, 0.95) within subjects ANOVA.

One participant was excluded from the analysis because they did not perform significantly above chance on the dot comparison task. This left 11 participants in the analysis.[1]

6.4 Results

As expected, there was a significant main effect of numerosity ratio on accuracy scores, $F(3, 30) = 12.27$, $p < .001$, $\eta_p^2 = .55$. Participants' accuracy decreased as ratios became closer to one (from 0.87 mean accuracy in the 0.7 ratio condition, to 0.62 mean accuracy in the 0.95 ratio condition). This represented a significant linear trend of numerosity ratio, $F(1, 10) = 37.05$, $p < .001$.

Importantly, there was a significant main effect of block type on accuracy scores, $F(1, 10) = 22.72$, $p = .001$, $\eta_p^2 = .69$. Participants' accuracy scores were significantly higher on incongruent test trials that were preceded by incongruent trials (block type 2, $M = .79$) than incongruent test trials that were preceded by congruent trials (block type 1, $M = .67$), $t(10) = -4.77$, $p < .001$ (See Figure 6.2)

There was no significant block type by ratio interaction, $F(3, 30) = 1.10$, $p = .36$.

6.5 Discussion

Recently, the hypothesis that comparing non-symbolic dot arrays may involve inhibitory control skills has developed from the finding that performance is superior on congruent in comparison to incongruent dot comparison task trials. Nevertheless, besides congruency effects, there is relatively

[1] Note that the same pattern of results emerged when all 12 participants were included in the analysis.

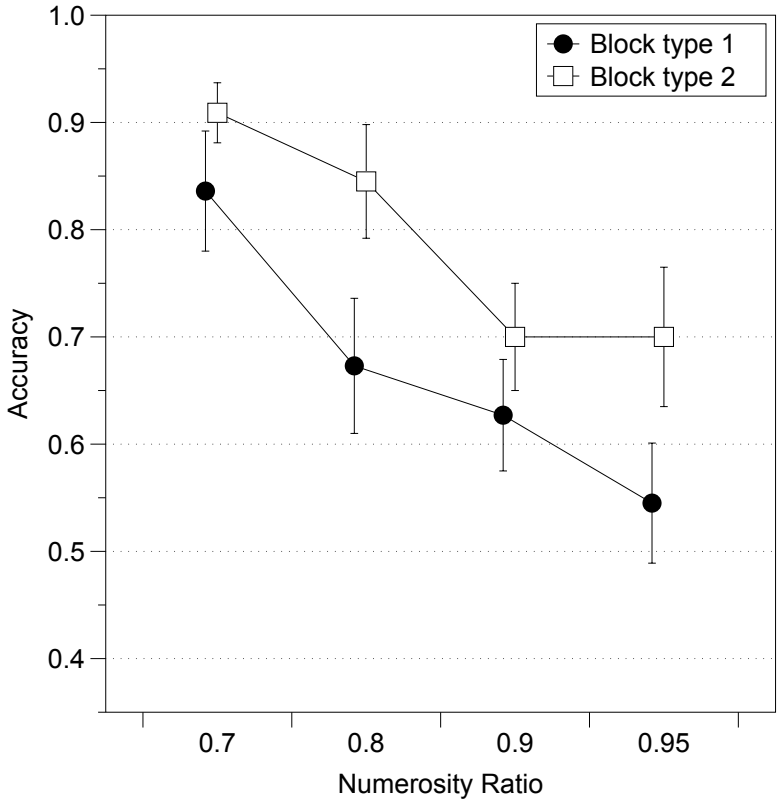


Figure 6.2: Mean accuracy scores on incongruent test trials either preceded by blocks of 5 congruent trials (block type 1), or blocks of 5 incongruent trials (block type 2). Error bars represent the standard error of the mean.

little evidence of a correlation between inhibition and dot comparison task performance (Cappelletti et al., 2014; Fuhs & McNeil, 2013) and conflicting evidence also exists (Keller & Libertus, 2015). In this study we showed that dot comparison task responses followed a pattern of performance comparable to that of classic inhibition tasks. The results indicated that the frequency of conflict of congruent and incongruent dot comparison trials affected performance in the same way that it has previously been shown to affect performance on Stroop task trials (Botvinick et al., 2001; Lindsay & Jacoby, 1994; Logan et al., 1984; Tzelgov et al., 1992). Specifically, accuracy on incongruent trials was significantly higher when preceded by several similarly incongruent trials, in comparison to when preceded by several congruent trials.

This result provides support for the competing processes hypothesis that inhibitory control is an important process involved in making dot comparison task judgements. Results are in line with Braver's (2012) theory: when incongruent test trials are preceded by multiple similarly incongruent trials, participants' expectations for interference to occur increases, and consequently proactive control is recruited. This means that the goal relevant information, in this case numerosity processing, is kept active in mind throughout. In comparison, when incongruent test trials are preceded by multiple congruent trials, the incongruent visual cues may cause an unexpected conflict, leading to lower accuracy.

Results from the present study contradict the Dehaene (1997) model which suggests that dot comparison task performance is only influenced by the individual's ANS acuity and the ratio difference between the dot arrays in each trial. Instead, results support the hypothesis that misleading visual characteristics may interfere with ANS processing during dot comparison tasks, and that inhibitory control skills are necessary to override this interference and respond correctly. In line with other inhibitory control task findings, the pattern of performance reported here suggests that processing is not completed on an independent, individual trial-by-trial basis, but that the expectation of conflict from previous trials in the task substantially influences responses (Botvinick et al., 2001; Lindsay & Jacoby, 1994; Logan et al., 1984; Tzelgov et al., 1992).

This finding holds important implications for studies that have found a relationship between dot comparison task performance and formal mathe-

matics achievement. Due to the commonly-reported relationship between inhibition skill and mathematics achievement (Blair & Razza, 2007; Bull & Scerif, 2001; Espy et al., 2004; St Clair-Thompson & Gathercole, 2001), it is critical that we understand whether inhibition also plays a role in dot comparison task performance. As the present results suggest that inhibition is recruited when completing incongruent dot comparison trials, these findings are in line with previous evidence demonstrating that the relationship between dot comparison task performance and mathematics achievement is mediated by inhibition ability (Fuhs & McNeil, 2013; Gilmore et al., 2013).

6.6 Summary of findings

This study shows that individuals' responses to dot comparison task trials follow a pattern of response similar to that of inhibition tasks. Specifically, participants performed more accurately on incongruent dot comparison trials when they were more frequent relative to congruent trials. This finding adds weight to the hypothesis that dot comparison tasks are not pure measures of the ANS, and inhibition is likely to be involved in making non-symbolic quantity judgements using this task. More evidence is needed, however, to demonstrate a direct link between individual differences in dot comparison task performance and inhibition task performance.

Chapter 7

Inhibition task correlations (Studies 4 and 5)

The results of the Studies 1, 2 and 3, presented in Chapters 3, 5 and 6 of this thesis, show a growing body of evidence to suggest that inhibition plays a meaningful role in dot comparison task performance. Study 1 showed that the visual characteristics of dot comparison stimuli have a substantial influence over task performance, with evidence that participants perform significantly less accurately on trials where the visual cues are incongruent, as opposed to congruent, with numerosity. Similarly, Study 2 replicated these visual cue congruency effects, and additionally showed that these effects varied with the set size of the dot arrays and the changing saliency of the visual cues. Finally, Study 3 demonstrated that dot comparison performance follows similar patterns to classic inhibitory control tasks, with incongruent trial accuracy scores dependent on the frequency of conflict between congruent and incongruent trials. Put together, these studies suggest that inhibition may be recruited to ignore misleading visual cues when responding to incongruent dot comparison task trials.

Nevertheless, the studies reported earlier in this thesis did not include an explicit measure of inhibitory control skills. A significant correlation between the magnitude of inhibition task congruency effects and the magnitude of dot comparison task congruency effects would provide convincing evidence of the role of inhibition in dot comparison task performance. Studies 4 and 5 reported in the current chapter aimed to demonstrate a direct link between dot comparison performance and classic inhibition task performance.

7.1 Introduction

Previous research investigating the relationship between dot comparison tasks and mathematics achievement has found that inhibition mediates this relationship (Fuhs & McNeil, 2013; Gilmore et al., 2013). However, few studies report the relationship between dot comparison and inhibition task scores.

Fuhs and McNeil (2013) found that individual differences in preschoolers' non-symbolic comparison accuracy scores were significantly correlated with their inhibitory control performance, as measured by overall accuracy on a composite of variations of the Day/Night task (say "day" when you see a picture of the moon, and "night" when you see a picture of the sun; similar Head/Feet and Knock/Tap tasks were administered). Fuhs and McNeil found evidence of this correlation with overall non-symbolic comparison task accuracy, as well as specifically with accuracy on trials where the cumulative surface area of the stimuli was not predictive of the numerosity represented (incongruent trials). Interestingly, the correlation between non-symbolic comparison accuracy and inhibition accuracy did not hold for trials where cumulative surface area was correlated with numerosity (congruent trials), presumably as there was no need to inhibit visual cues. This finding strongly supports the hypothesis that inhibition is pivotal to performance on incongruent non-symbolic comparison task trials. However, the stimuli used in this study were stars rather than dots and, given how sensitive individuals' performances on non-symbolic comparison tasks are to different visual cue characteristics (see Study 1), it is possible that results are not entirely generalisable to more standard dot comparison tasks.

In line with Fuhs and McNeil's results, Cappelletti et al. (2014) found that older participants' (aged 60–75) dot comparison w scores on incongruent trials correlated with their reaction time performance on the incongruent trials of two different Stroop tasks (a number Stroop and a word Stroop). Conversely, there was no significant correlation between participants' dot comparison w scores on the congruent trials and either measure of inhibition. In addition, the equivalent analyses with younger adult participants (19–36 years) in Cappelletti et al.'s study revealed no relationship between dot comparison task w scores on congruent or incongruent trials and either of the inhibition task measures. Older adults tend to have impoverished

inhibitory control skills in comparison to younger adults (Cappelletti et al., 2014; Hasher, Zacks, & May, 1999), and as such demonstrated larger congruency effects (response time differences between congruent and incongruent trials) than younger adults on both the dot comparison task and the Stroop tasks (Cappelletti et al., 2014). In contrast, the younger adults did not actually show a congruency effect on the dot comparison task. Their performance on congruent trials was not significantly different to their performance on incongruent trials, which might account for the lack of correlation with inhibition tasks scores. Importantly, Cappelletti et al.'s dot comparison stimuli were created without any controls for convex-hull size, a cue found to be particularly influential, and therefore it is not entirely surprising that the younger adults were not influenced by the partial controls for visual cue confounds (see Study 1, Chapter 3 for evidence relating to this finding). It is possible that older adults with weaker inhibitory control skills were more sensitive to the minimal inhibition demands of this dot comparison task.

Nevertheless, a similar non-significant correlation was also recently reported by Keller and Libertus (2015) in a study of 5–6 year old children's dot comparison performance. This study found no differences between congruency conditions (again, the stimuli did not control for convex-hull size), and reported no significant correlation between dot comparison task accuracy and performance on a the NEPSY-II inhibition subtest, a measure of the interference control sub-type of inhibition. But it did find a significant correlation between dot comparison task accuracy and mathematics achievement. In contrast, Gilmore et al. (2013) used the same inhibition task (from the NEPSY-II) with 7–10 year old children and found that inhibition task performance mediated the relationship between dot comparison task performance and mathematics achievement. This finding conflicts with that of Keller and Libertus (2015) and suggests that inhibition may in fact be involved in dot comparison task performance. These conflicting findings may relate to methodological differences in the dot comparison and inhibition tasks used in the two studies, as discussed previously in Chapter 4, Section 4.5.

The findings summarised above provide mixed evidence of whether a correlation between dot comparison task performance and inhibition ability exists. Discrepancies are likely due to differences in the method of creating dot comparison task stimuli, specifically divergent controls for visual cues.

Additionally the age of the participants may cause some disparities in results, with limited evidence from only a handful of studies across a wide range of development ranging from preschoolers to older adults. Therefore, the role of inhibition in non-symbolic comparison remains unclear.

The two studies reported here aimed to investigate the relationship between individual differences in dot comparison task performance and individual differences in inhibition task performance. The first study, Study 4, is an analysis of further inhibition task data that was collected alongside the dot comparison task of Study 1 (Chapter 3), which investigated the influence of different protocols for controlling visual cue stimuli. For the purposes of this investigation, data from the dot comparison task trials created with the Gebuis and Reynvoet (2011) protocol is analysed alongside participants' performances on an interference control inhibition task, the colour-word Stroop task, that was performed concurrently. To explore how variations in inhibition task procedures influenced the relationship with dot comparison performance, the second study reported in this chapter, Study 5, investigated dot comparison performance alongside three different interference control inhibition tasks, including the Flanker task and two modified Stroop tasks. Both adults' and children's performance was investigated in Study 5 to explore whether the relationship between dot comparison performance and inhibition changes as inhibition skills develop.

7.2 Study 4

7.2.1 Method

The method of the dot comparison task used in this study was reported in Study 1 (Chapter 3, Section 3.2), therefore I provide an overview here, alongside additional details of the inhibitory control task that was not previously described in Study 1.

7.2.1.1 Participants

Participants were 57 adult students from Loughborough University (24 male, 33 female) with a mean age of 21.34 years ($SD = 2.35$). Participants were tested individually in a quiet room and were given a £3 inconvenience allowance for their time. This study was approved by the Loughborough

University Ethics Approvals (Human Participants) Sub-Committee.

7.2.1.2 Tasks

All participants completed two tasks on a computer: a dot comparison task and a colour-word Stroop task. Tasks were presented in a counterbalanced order.

7.2.1.3 Dot comparison task

Participants briefly viewed two arrays of dots on a screen and were required to select the more numerous array. The two arrays consisted of blue or yellow dots on a grey background and were presented simultaneously, side-by-side on a 15" laptop screen. Participants were asked to select which array was more numerous using left and right keys marked on the keyboard. Each trial began with a fixation point (600 ms) followed by presentation of the two arrays (600 ms) and finally a grey screen with a white '?' was presented in the centre until a response was given. The task took approximately 15 minutes to complete.

The trials consisted of two types of dot comparison stimuli: arrays created using the Gebuis and Reynvoet (2011) protocol, and arrays created using Panamath software (identical stimuli to those used by Libertus et al., 2012). There were eight practice trials followed by a total of 312 experimental trials, which were divided into four blocks (for further details see Section 3.2.2). For the purposes of this investigation, only trials from block one (96 Gebuis and Reynvoet protocol trials) were analysed. The analyses in Chapter 3, Study 1, showed that participants' accuracy scores on trials created using these divergent protocols were not significantly correlated, and concluded that these different protocol trials do not appear to be measuring the same cognitive construct. Consequently, the Gebuis and Reynvoet protocol trials were selected for this analysis due to the extremely low reliability of the Panamath trials. The Gebuis and Reynvoet trials comprised four image types: fully congruent, dot-size congruent and convex-hull incongruent, dot-size incongruent and convex-hull congruent, and fully incongruent (for further details see Chapter 3, Section 3.2.3). Trial set size ranged from 22 to 36 dots and numerosity ratios ranged between 0.61 and 1.64.

7.2.1.4 Colour-word Stroop task

Participants completed a colour-word version of the Stroop task (Stroop, 1935) presented on a 15" laptop screen. This involved responding to a written colour word presented in the centre of the screen, whilst ignoring the font colour of the text. Participants completed 40 trials in total, split into two blocks of 20 trials. Block one included congruent trials, where the font colour matched the written word (e.g. BLUE, GREEN, RED). Block two included incongruent trials, where the font colour did not match the written word (e.g. BLUE, GREEN, RED). The order of the blocks was counterbalanced. Participants responded by pressing a coloured key on the keyboard that corresponded to the written word on screen, as quickly and as accurately as they could. The task took under 5 minutes to complete.

7.2.2 Analysis

Performance on the dot comparison task was measured with mean accuracy scores. Accuracy on the Stroop task was close to ceiling, and so performance was measured with median response times (RT) for trials answered correctly.

The influence of trial congruency on each task was analysed using paired-samples t-tests to examine whether performance was significantly different on congruent trials and incongruent trials. Congruency effects were then calculated for the Stroop task using response time differences (incongruent trial RT − congruent trial RT) as a measure of inhibition, with a smaller difference indicating better inhibition skill. Congruency effects for the dot comparison task were similarly calculated using accuracy on the fully congruent trials − accuracy on the fully incongruent trials. Pearson correlations were conducted to investigate whether there was a significant relationship between congruency effects on the Stroop task and congruency effects on the dot comparison task.

Six participants were excluded from the analysis because English was not their first language. For the Stoop task to measure inhibitory control, it is necessary that the words are processed automatically (Nigg, 2000). For participants who did not speak English as a first language, reading the words may have been a more effortful process and therefore easier to ignore when required by the demands of the task (MacLeod, 1991). Thus, performance for these participants may not have reflected their inhibitory control skills

to the same extent as native English speakers. One further participant was excluded for misunderstanding the Stroop task instructions, responding based on font colour rather than the written word, resulting in 0% accuracy on the incongruent trials. Finally, five participants were excluded from the analysis because they did not perform significantly above chance on the dot comparison task trials. This left 45 participants in the analysis.

7.2.3 Results

7.2.3.1 Congruency effects

Participants demonstrated significant congruency effects on both tasks. Firstly, participants performed significantly more accurately on fully congruent dot comparison trials ($M = 0.90$ accuracy, $SD = 0.11$) than fully incongruent dot comparison trials ($M = 0.47$ accuracy, $SD = 0.19$), $t(45) = 11.01$, $p < .001$. Participants also performed significantly more accurately on convex-hull congruent, dot-size incongruent trials ($M = 0.90$ accuracy, $SD = 0.10$), than convex-hull incongruent, dot-size congruent trials ($M = 0.63$ accuracy, $SD = 0.19$), $t(45) = 6.82$, $p < .001$.

Similarly, participants performed significantly faster on congruent Stroop task trials ($Mdn = 790$ milliseconds, $SD = 100$), than incongruent trials ($Mdn = 1010$ milliseconds, $SD = 190$), $t(45) = 9.53$, $p < .001$.

7.2.3.2 Correlations between tasks

Pearson correlations were conducted to investigate whether there was a significant relationship between participants' congruency effects on the dot comparison task, and their congruency effects on the Stroop task. However, results showed no significant correlation between dot comparison congruency effects (fully congruent accuracy score − fully incongruent accuracy score) and Stroop task congruency effects (incongruent RT − congruent RT), $r = -.143$ $p = .349$.

7.2.4 Discussion

This study aimed to investigate whether inhibition skills were recruited during dot comparison tasks by exploring the relationship between participants' dot comparison performance and their performance on a Stroop inhibition

task. The present results showed that although participants' performances were significantly influenced by the congruency status of the trials in both tasks, i.e. better performance on the congruent in comparison to incongruent trials, there was no significant correlation between these congruency effects across the two tasks. Given the substantial differences in performance between congruent and incongruent trials on both tasks (indicating inhibition was involved in both tasks) this finding was unexpected.

Research has shown that there are many different types of inhibition (Nigg, 2000), and so it is possible that performance on the dot comparison task and performance on the Stroop task require different types of inhibitory control skill. However, the colour-word Stroop task was selected as a classic and widely-used measure of interference control inhibition, the inhibition sub-type defined by Nigg (2000) as the ability to maintain a primary response in the presence of distracting stimuli pulling for a competing response. This seems fitting to the hypothesis that inhibition has a role in incongruent dot comparison task performance because successful performance depends on the ability to respond to the numerosity of the arrays, whilst ignoring distracting and misleading visual cue information, such as the size of the dots or the convex hull. The finding that there is no correlation between congruency effects from these two tasks with purportedly similar inhibitory demands suggests that there are nuances in the task format that lead to divergent cognitive processing.

Indeed, a study by Shilling, Chetwynd, and Rabbitt (2002) found that differences in the task demands of interference control tasks have a substantial influence on the level of individual consistency in performance across the measures. Shilling et al. investigated older adults' performance on four analogues of the traditional Stroop task. The first of these was a traditional colour-word Stroop task. The second was a 'figure ground' task, where the aim was to respond based on the individual digits that combined to make up a larger 'global digit', e.g. many 3s making up the shape of a larger 8. The third interference task required participants to respond to the direction of arrows on a screen, whilst inhibiting the written direction word that appeared inside the arrow, e.g. 'right' written inside an arrow that pointed left. The final task was a number modification of the Stroop task where the task was to respond based on the total number of arabic numerals presented on the screen, and ignore the identity of the digits themselves, e.g. for the stim-

uli '3333' the correct answer would be 'four'. Shilling et al. (2002) reported no evidence that individuals who were particularly sensitive to interference on one measure (i.e. showed large congruency effects), were also sensitive to interference on the other analogues of the task. Specifically, in their model of performance, the estimated correlations between individual differences in congruency effects across the tasks were very weak (all non-significant, rs $<$ .244).

In a follow-up study, Shilling et al. (2002) demonstrated that relationships between individual differences in performance across multiple inhibition tasks were improved by increasing the similarity of the surface demands of the task. Here, the authors showed that individuals' performances on two variations of the arrows task, described above, correlated when the only difference between the tasks was that one version used up and down arrows, and the other used left and right arrows. This may seem an obvious finding, but this study provides evidence that performances on inhibition tasks assumed to measure the same sub-type of inhibition (interference control), do not correlate unless the demands of the task are extremely similar.

To investigate this finding further with regard to the relationship between dot comparison congruency effects and inhibition congruency effects, the following study explored dot comparison task performance alongside three additional interference control inhibition tasks, with methodological formats that are more akin to the dot comparison task. Specifically, the following study explored dot comparison performance in a new cohort of participants using exactly the same trials reported in Study 4 above, alongside three different inhibition measures: an animal size Stroop task, a number size Stroop task, and a Flanker task (described in detail in Section 7.3.1.2. The animal size and number size Stroop tasks are variations on the original colour-word Stroop task (Stroop, 1935), except that the response options consist of two sets of stimuli presented simultaneously on screen. Thus, these tasks were chosen because the presentation and response format is more similar to the dot comparison task. Both the animal and number versions of these tasks were administered to explore whether there were any differences in the relationships with dot comparison task performance due to the numerical processing aspect of the number size Stroop task. Finally, the Flanker task was administered as a classic measure of interference control inhibition that is frequently used within the inhibition literature. Unlike many other

inhibition tasks, and similarly to the dot comparison task, the Flanker task requires minimal real-world knowledge (e.g. knowledge of animals, numerical order of arabic digits, word reading) to respond. To investigate whether the relationships between these tasks changes with the development of inhibition skills, both adults, with supposedly fully developed inhibition skills, and 7–11 year old children, still developing their inhibition abilities, participated in this study (Nigg, 2000).

It was predicted that due to the greater similarities between the methodological formats of the three inhibition tasks and the dot comparison task, significant correlations between congruency effects across the tasks may be demonstrated.

7.3 Study 5

7.3.1 Method

7.3.1.1 Participants

Participants were 51 adult students from Loughborough University (19 Male, 32 Female), with a mean age of 24.47 years (SD = 4.50) and 80 children aged 7–11 years (42 Male, 38 female), with a mean age of 9.5 (SD = 1.27) Participants were tested individually in a quiet room. Adults were given a £4 inconvenience allowance for their time, and children received game tokens as part of a Summer Scientist Week event (www.summerscientist.org). This study was approved by the Loughborough University Ethics Approvals (Human Participants) Sub-Committee and the University of Nottingham ethics committee.

7.3.1.2 Tasks

All participants completed three tasks on a computer: a dot comparison task, an animal size Stroop task, and a number size Stroop task. Adults additionally completed a Flanker task. The children did not complete this task due to restrictions on testing time for the Summer Scientist Week event. Further details of each of the tasks are presented in turn below. Tasks were presented in a counterbalanced order.

7.3.1.3 Dot comparison task

The dot comparison task trials were identical to those used for the analysis in the above study (Section 7.2.1.3). In contrast to the procedure of Study 1, participants only completed these 96 trials, and did not complete any further dot comparison trials in the battery of tasks.

7.3.1.4 Animal size Stroop task

The animal size Stroop task (based on the animal size Stroop task reported in Szűcs, Devine, Soltesz, Nobes, & Gabriel, 2013), was designed to assess participants' ability to inhibit irrelevant information in a non-numerical context. Participants viewed two pictures of animals on a screen and were required to select the larger animal in real life as quickly as possible (Figure 7.1). One animal was selected from a set of large animals (e.g. a bear, gorilla, or giraffe), and the other was selected from a set of small animals (e.g. an ant, rabbit, or mouse). One animal image was presented four times larger in area than the other animal image. On congruent trials, the larger animal on-screen was also larger in real life. On incongruent trials, the larger animal on-screen was smaller in real life. The task was made up of 50% congruent trials and 50% incongruent trials presented in a random order. Images were presented simultaneously on screen until the participant responded. Participants responded by pressing the left and right keys marked on the keyboard corresponding to each side of the screen. The task included 8 practice trials and 96 experimental trials. In order to ensure that participants had the necessary real-world knowledge to complete the task, participants were shown pictures of each animal prior to commencing the task, and asked whether the animal was large or small in real life. The task took under 5 minutes to complete.

7.3.1.5 Number size Stroop task

The number size Stroop task followed the same procedures as the animal size Stroop task described above, except the stimuli were Arabic numerals instead of animals. Therefore, on a congruent trial the numerically larger number was presented four times as large on screen as the numerically smaller number. On an incongruent trial the numerically larger number was presented four times smaller on screen as the numerically smaller number

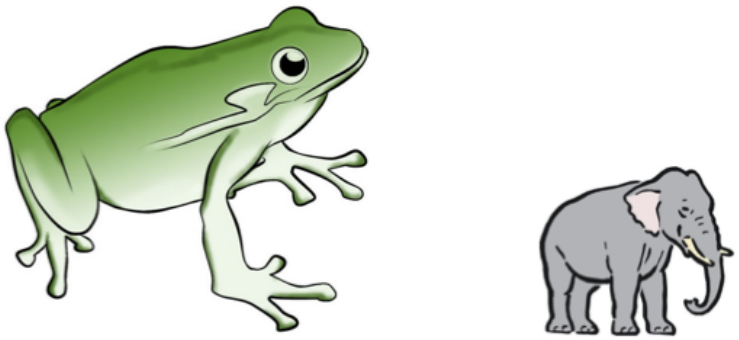


Figure 7.1: An example of an animal size Stroop task trial (incongruent trial).

(Figure 7.2).

Figure 7.2: An example of a number size Stroop task trial (incongruent trial).

7.3.1.6 Flanker task

The Flanker task was included as a standard measure of interference control, using non-numerical stimuli that did not require any real-world knowledge. During this task participants viewed a row of five arrows on screen and were required to select the direction the middle arrow was pointing, whilst

ignoring the direction of the flanking arrows around the outside (Figure 7.3). The flanking arrows could either be pointing in the same direction as the central arrow (congruent trials), or in the opposite direction (incongruent trials). The task was made up of 50% congruent trials and 50% incongruent trials, presented in a random order. The stimuli were presented on screen until the participant responded by pressing the left and right keys marked on the keyboard. The task included 8 practice trials and 80 experimental trials. The task took under 5 minutes to complete.

Figure 7.3: An example of a Flanker task trial (incongruent trial).

7.3.2 Analysis

Performance on the dot comparison task was measured with mean accuracy scores. Because accuracy scores for each of the inhibition tasks (animal size Stroop, number size Stroop and Flanker) were close to ceiling, median response times were used for trials answered correctly.

Congruency effects for individual tasks, and correlations between each of the tasks were analysed as described in the previous study (Section 7.2.2), using response time differences for the inhibition measures, and accuracy differences for the dot comparison task.

Seventeen participants (16 children and one adult) were excluded from the analyses because they did not perform significantly above chance on the dot comparison task. Additionally, one adult participant was excluded from the cross-task correlations involving the Flanker task, because they did not complete the full number of trials for this task.

As a preliminary analysis, to assess whether data from the children and adult groups should be analysed separately, two regressions were run with dot comparison congruency effect as the dependent measure. In the first, the predictors were group (adult or child), number Stroop congruency, and

the group by number Stroop congruency interaction. In the second the predictors were group, animal Stroop congruency and the group by animal Stroop congruency interaction. In neither case did the interaction effects approach significance, ps = .704, .190 respectively. Given this, the adult and child data were not separated for the main analysis.

7.3.3 Results

7.3.3.1 Congruency effects

Participants demonstrated significant congruency effects for all four tasks (see Figure 7.4). Firstly, participants performed significantly more accurately on fully congruent dot comparison trials (M = 0.89 accuracy, SD = 0.12) than fully incongruent dot comparison trials (M = 0.47 accuracy, SD = 0.21), $t(113) = 15.54$, $p < .001$. Participants also performed significantly more accurately on convex-hull congruent, dot-size incongruent trials (M = 0.79 accuracy, SD = 0.18), than convex-hull incongruent, dot-size congruent trials (M = 0.69 accuracy, SD = 0.21), $t(113) = 2.95$, $p = .004$.

Participants performed significantly faster on congruent animal Stroop task trials (Mdn = 710 milliseconds, SD = 170), than incongruent trials (Mdn = 830 milliseconds, SD = 230), $t(113) = 13.58$, $p < .001$.

Participants performed significantly faster on congruent number Stroop task trials (Mdn = 800 milliseconds, SD = 190), than incongruent trials (Mdn = 880 milliseconds, SD = 230), $t(113) = 13.34$, $p < .001$.

Participants performed significantly faster on congruent Flanker task trials (Mdn = 510 milliseconds, SD = 60), than incongruent trials (Mdn = 570 milliseconds, SD = 70), $t(49) = 14.11$, $p < .001$.

7.3.3.2 Correlations between tasks

Pearson correlations were conducted to investigate whether there was a relationship between participants' performances on the different inhibition measures (see Table 7.1). There was a significant correlation between congruency effects on the animal Stroop and number Stroop tasks ($r = .498$, $p < .001$), but no significant correlations between performances on the animal Stroop and the Flanker task ($r = .231$, $p = .106$), and the number Stroop and the

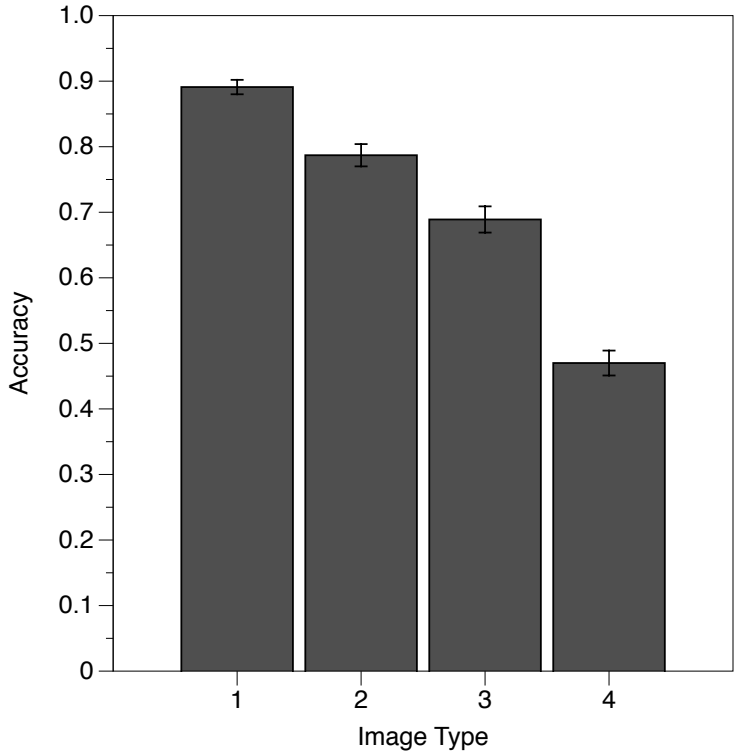


Figure 7.4: Mean dot comparison task accuracy scores for each image type. Image type 1 represents fully congruent trials (convex-hull congruent and dot-size congruent); Image type 2 represents convex-hull congruent, dot-size incongruent trials; Image type 3 represents convex-hull incongruent, dot-size congruent trials; Image type 4 represents fully incongruent trials (convex-hull incongruent and dot-size incongruent). Error bars represent the standard error of the mean.

Flanker task ($r = .255$, $p = .074$).[1]

Pearson correlations were used to investigate whether individuals who showed a smaller congruency effect on the inhibition tasks also showed a smaller congruency effect on the dot comparison task (see Table 7.1). There was a significant correlation between dot comparison congruency effects (fully congruent accuracy score − fully incongruent accuracy score) and animal Stroop congruency effects (incongruent RT − congruent RT; $r = .227$, $p = .015$), as well as number Stroop congruency effects (incongruent RT − congruent RT; $r = .198$, $p = .035$). There was no significant correlation found between dot comparison congruency effects and Flanker task congruency effects (incongruent RT − congruent RT; $r = .090$, $p = .534$).

	1	2	3	4
1. Flanker	-			
2. Animal size Stroop	.231	-		
3. Number size Stroop	.255	.498**	-	
4. Dot comparison	.090	.227*	.198*	-

Note. *$p < .05$, **$p < .001$.

Table 7.1: Pearson correlation coefficients for congruency effects on each task

7.3.4 Discussion

This study aimed to further investigate the findings from the Study 4 Stroop task analysis which found no significant correlation between dot comparison task congruency effects and colour-word Stroop task congruency effects. In this follow-up study, participants completed a dot comparison task alongside three different interference control tasks, with task requirements that were more closely related to those of the dot comparison task.

First, in line with previous findings from Study 4 (Section 7.2.3.1) this study showed that performances on all three inhibition tasks and the dot comparison task were influenced by the congruency status of the task trials. Participants performed significantly more accurately on the fully congru-

[1]Note that only adults completed the Flanker task.

ent dot comparison task trials in comparison to the fully incongruent dot comparison task trials, and performed significantly faster on all congruent inhibition task trials than on incongruent inhibition task trials.

Second, this study demonstrated that dot comparison task performance does correlate with inhibition task performance when the task format is very similar to the dot comparison task. Specifically, although there was no significant correlation between dot comparison congruency effects and one measure of interference control, the Flanker task, there were significant correlations with congruency effects obtained in the animal size and number size Stroop tasks. The animal size and number size Stroop tasks were presented in a similar format to the dot comparison task. On each of these three tasks the stimuli included two distinct images where the physical size of the object was irrelevant to the goals of the task. These stimuli were presented simultaneously, side-by-side, and participants were required to respond to the side of the screen that contained the semantically larger or more numerous stimuli. In comparison, the methodological format of the Flanker task differed somewhat, with the task-irrelevant information consisting of arrows with unhelpful semantic value, rather than unhelpful physical size. In the Flanker task participants were required to focus on a single part of one central image and ignore the surrounding irrelevant information, in contrast to weighing up two choices in visually distinct areas of the display.

Although these differences between task formats may seem negligible, previous research in the inhibition domain has shown that performances on inhibition tasks supposedly measuring the same sub-type (e.g. interference control), do not always correlate (Shilling et al., 2002). It is likely that the cognitive mechanisms underlying inhibitory control differ depending on nuances in task methodologies. Consequently it is possible that there are multiple different types of inhibition that are yet to be distinguished or categorised.

7.4 Summary of findings (Study 4 and Study 5)

To summarise, the results reported in this chapter have provided evidence of the role of inhibitory control in dot comparison task performance. Inhibition task congruency effects were found to be significantly correlated with dot comparison task congruency effects when task formats were similar. In

contrast, when inhibition tasks involved different stimuli presentation or response options, the correlations were non-significant. This finding is in line with previous research showing a lack of consistency in individual differences in performance on multiple versions of interference control tasks (Shilling et al., 2002). Combined with the data from Study 2 and Study 3 (Chapters 5 and 6) demonstrating dot comparison performance patterns commensurate with inhibition task data, there is now a substantial body of evidence to suggest that inhibition skills are involved in completing incongruent dot comparison task trials.

Part IV

Do non-symbolic numerosity tasks involve numerosity processing?

Chapter 8

Developmental differences in the use of numerosity and visual cues

To recap, Part II of this thesis reported empirical evidence that the visual cues in dot comparison stimuli have a significant influence on task performance. In fact, differences between stimuli generation methods influenced individuals' judgements so substantially that the same participants' performances on two variations of the task were found to be statistically unrelated. Part III of this thesis demonstrated that individual differences in dot comparison task accuracy can, in part, be explained by individual differences in inhibitory control skills. Put together, these findings suggest that dot comparison tasks are not pure measures of ANS acuity. In contrast, the results described above lead to the question of how much numerosity processing plays a role in dot comparison task performance, if at all. The present study brings together findings from three dot comparison tasks reported in this thesis to examine whether ANS processing influences task performance over and above visual cue processing. The results are discussed with relevance to the future use of dot comparison tasks as measures of ANS acuity.

8.1 Introduction

Recent research has shown that dot comparison tasks do not exclusively measure ANS acuity, and that the visual characteristics of the dot array stimuli

also substantially influence judgements (Fuhs & McNeil, 2013; Gebuis & Reynvoet, 2012a; Gilmore et al., 2013; Smets et al., 2015; Szűcs et al., 2015). Notably, Gebuis and Reynvoet (2012a) found that individuals weigh up and integrate information from multiple visual cues in order to make judgements of numerosity. From this, Gebuis and Reynvoet concluded that the existence of an ANS that is independent of visual cues appears unlikely. Others have suggested that dot comparison task performance may be influenced by both ANS acuity and other competing processes, such as inhibition, that are driven by visual cue processing (Cappelletti et al., 2014; Fuhs & McNeil, 2013; Gilmore et al., 2013; Nys & Content, 2012; Szűcs et al., 2015). Finally, the traditional model of the ANS (Barth et al., 2005; Dehaene, 1997) assumes dot comparison performance is influenced by the acuity of an individual's ANS representation, independent of the visual characteristics of the task (Feigenson et al., 2004).

The present study aimed to explore which of the above theories most accurately describes dot comparison task performance for children and for adults. It has already been established that the traditional model of the ANS cannot explain many of the patterns of results reported in this thesis, from the influence of visual cue controls (Study 1), to set size effects (Study 2), to frequency of conflict congruency effects (Study 3). Therefore, the question remains whether visual cue processing can entirely account for dot comparison task performance, without any additional influence of ANS processing, in accordance with Gebuis and Reynvoet's (2012a) suggestion.

In order to explore this question systematically with a large sample, dot comparison task performances from three studies reported in this thesis were re-analysed. Only standard tasks comprising of equal proportions of congruent and incongruent trials were included, therefore the dot comparison task reported in Study 3 (Chapter 6), used to investigate the influence of changing proportions of congruent vs. incongruent trials, was not included. Nevertheless, data from Studies 1, 2, and 5 (Study 4 reported the same dot comparison data as Study 1) were collated to form a new data set for this re-analysis.

The aim of this study was to investigate whether participants used numerosity information over and above the stimuli's visual cues when comparing dot arrays, and whether findings were consistent across multiple dot comparison tasks, including tasks created with different controls for visual

cues. There were three main research questions. First, is numerosity information predictive of dot comparison task accuracy scores, after controlling for visual cue information? Second, are there any developmental differences in this relationship? Finally, are there significant differences across dot comparison tasks using different stimuli? The answers to these questions will help to assess the validity of dot comparison tasks as a measure of the ANS acuity of adults and children.

8.2 Method

A total of 244 participants (124 children, mean age = 9.19, $SD = 1.25$; 120 adults, mean age = 22.86, $SD = 3.85$) completed dot comparison tasks in three separate studies. The three studies are reported separately in this thesis in full, so only a brief reminder of the participants and methodologies of each is provided here.

8.2.1 Study 1 overview

Participants were 57 adult students from Loughborough University (24 male, 33 female) with a mean age of 21.34 years ($SD = 2.35$). The dot comparison task included 120 trials created with the Panamath software (Halberda et al., 2008), and 192 trials created with the Gebuis and Reynvoet (2011) protocol. In total, participants completed 312 experimental trials.

8.2.2 Study 2 overview

Participants were 44 children (21 male, 23 female) aged 7–9 years ($M = 8.36$, $SD = 0.60$ years), and 12 adults (3 male, 9 female) aged between 19 and 31 years ($M = 23.20$, $SD = 4.04$ years). The dot comparison task included 184 trials created with the Gebuis and Reynvoet (2011) protocol.

8.2.3 Study 5 overview

Participants were 51 adult students from Loughborough University (19 male, 32 female), with a mean age of 24.47 years ($SD = 4.50$), and 80 children aged 7–11 years (42 male, 38 female), with a mean age of 9.65 years ($SD = 1.27$). The dot comparison task included 96 trials created with the Gebuis and Reynvoet (2011) protocol.

8.3 Analysis

For each of the stimuli used in the three experiments, the average dot size and the convex-hull size of each array was calculated. The Graham Scan algorithm (Graham, 1972) was used to calculate the size of the convex hull as described in Chapter 3, Section 3.3.1; average dot size was calculated by summing the number of coloured pixels in each array and dividing by the number of dots. Using these values, the ratio differences between the two arrays comprising each trial were calculated in terms of convex-hull size, average dot size and numerosity. These ratios were log transformed to produce a linear scale.

Each participant's trial-by-trial accuracy scores were subjected to separate hierarchical logistic regressions, predicting accuracy for every trial with two steps: step one included dot-size ratio and convex-hull size ratio (visual cues), step two included numerosity ratio. The change in pseudo R^2 values from the addition of step two was recorded. Additionally, whether or not numerosity information significantly independently predicted accuracy scores in step two of the regression was recorded as binary data (either significant or non-significant). This analysis aimed to capture whether, for each participant individually, accuracy on the dot comparison trials was significantly predicted by numerosity ratio after visual cue information was taken into account.

These data from all 244 participants were combined across studies 1, 2 and 5. Two sets of analyses were conducted involving different sets of trials. First, only trials created with the Gebuis and Reynvoet method were considered because previous research has demonstrated a non-significant correlation between performance on tasks created with different controls for visual cues (Smets et al., 2015; Study 1 findings). For this analysis, data from all 244 participants were combined across the three studies. A Mann-Whitney U test was used to compare differences in adults' and childrens' pseudo R^2 increase due to the inclusion of numerosity in the model. Pearson's chi square tests were then used to examine whether adults and children differed in their use of numerosity information, as measured by whether or not numerosity information significantly predicted their individual accuracy scores in step two of the regression model, after visual cues were controlled for. Chi square tests were also used to examine whether there was any dif-

ference in use of numerosity information across the three studies (Study 1, 2 and 5 as described above).

Second, data from the 57 participants in Study 1 were examined to explore whether different protocols had any influence on adults' use of numerosity information independent from visual cues. For this analysis the hierarchical logistic regression was performed twice for each participant, once with the Gebuis and Reynvoet protocol trials, and once with the Panamath protocol trails. A McNemar test was used to compare differences in whether numerosity information significantly predicted participants' accuracy scores, over and above visual cues, between trials created with each protocol.

No participants were excluded from these analyses. The aim of this study was to assess the relative influence of visual cues and numerosity processing in dot comparison tasks, so the decision to include participants who did not perform significantly above chance was made so that participants who were particularly reliant on visual characteristics were included.

8.4 Results

For each study, the changes in pseudo R^2 values, and the percentage of participants for whom numerosity ratio significantly predicted accuracy scores after controlling for visual cues, are presented in Table 8.1. Figure 8.1 shows that adults demonstrated larger increases in pseudo R^2 values due to the addition of numerosity information in the model at step two, when controlling for visual cue information in step one. This increase in pseudo R^2 values for the adults ($Mdn = 0.059$) represented a significantly larger increase in comparison to the change in children's pseudo R^2 values ($Mdn = 0.015$), $U = 3819$, $p < .001$. In line with this, a chi-square test of independence showed a significant effect of age group on whether or not numerosity information significantly independently predicted accuracy scores in step two of the regression, χ^2 $(1, N = 244) = 37.78$, $p < .001$, $\Phi = .39$. The addition of numerosity information to the model explained significantly more variance in accuracy scores than visual cues alone for 70.0% of adults, and just 30.6% of children. This means that for a majority of children (69.4%), and a large minority of adults (30%) accuracy on dot comparison trials could be accounted for without the need to include numerosity information in the model.

	Children		Adults	
	Median R^2 change	% sig	Median R^2 change	% sig
Study 1	-	-	0.045	75.4%
Study 2	0.013	36.4%	0.094	83.3%
Study 5	0.017	27.5%	0.065	60.8%
Total	0.015	30.6%	0.059	70.0%

Table 8.1: The median pseudo R^2 change when numerosity ratio was added to the regression models and the percentage of participants for whom numerosity ratio significantly predicted accuracy scores after controlling for visual cues, across all three experiments. Data from the Panamath trials was not included here.

This effect was consistent across multiple dot comparison studies for both adults and children. Chi-square tests of independence showed no effect of study on whether numerosity information significantly predicted participant's accuracy scores in step two of the model, when controlling for visual cues at step one. The effects of study characteristics were non-significant for children, χ^2 (1, $N = 214$) = 1.05, $p = .306$, and adults χ^2 (2, $N = 120$) = 3.88, $p = .144$.

A final analysis was conducted with the data from Study 1 to explore the influence of the protocol used to construct dot array stimuli (i.e. Gebuis & Reynvoet; Panamath). A McNemar test demonstrated that the method of stimuli construction had no significant effect on whether or not participants' accuracy scores could be significantly predicted by numerosity information over and above visual cues, $p = .815$. Adult participants were just as likely to use numerosity information over and above visual cue information on trials created with the Gebuis and Reynvoet protocol (75.4%), as on trials created with the Panamath protocol (71.9%).

8.5 Discussion

Non-symbolic dot comparison tasks are assumed to measure ANS acuity, but very few studies have explored the validity of this widely used task. Recently, evidence has highlighted the significant influence of visual cue processing on dot comparison performance (Fuhs & McNeil, 2013; Gebuis &

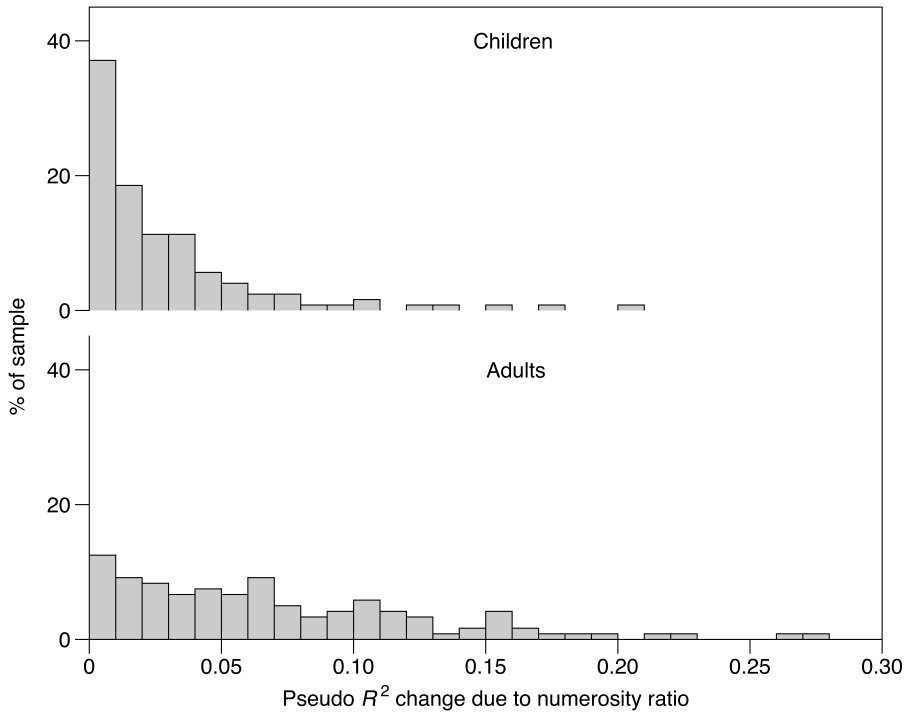


Figure 8.1: Change in pseudo R^2 values when numerosity ratio was added to regression models individually predicting accuracy scores for children (top panel) and adults (bottom panel). Data from the Panamath trials was not included here.

Reynvoet, 2012a; Gilmore et al., 2013; Smets et al., 2015; Szűcs et al., 2015). This study investigated whether both the visual characteristics (specifically convex hull and average dot size) and the numerical characteristics (i.e. the number of dots in the array) of dot comparison trials influenced accuracy scores in a large sample of children and adults. Results demonstrated that for the majority of children, numerosity information did not significantly explain any additional variance in accuracy scores over and above visual cue information. For most adults, however, numerosity information was predictive of accuracy scores even when controlling for visual cue information. There were no significant differences in these findings across the three experiments, or between trials created with different visual cue controls.

These findings have several implications for the use of dot comparison tasks in research intending to assess ANS acuity. First, and most crucially, the current findings suggest that dot comparison tasks may not be suitable as a measure of ANS acuity for all children. For almost 70% of children, numerical judgements did not explain significant extra variance in accuracy scores over and above that explained by visual cues. This has serious implications for studies that have investigated the correlation between dot comparison performance and symbolic mathematics achievement. Conclusions about this relationship that are based on dot comparison performance as a measure of ANS acuity may be invalid. This is particularly important because a large proportion of the studies investigating the acuity of numerical representations that may underlie mathematical achievement have been conducted with school-aged children (e.g. Bonny & Lourenco, 2013; Gilmore et al., 2013; Holloway & Ansari, 2009; Libertus et al., 2011; Nosworthy et al., 2013; Sasanguie et al., 2013; Wei et al., 2012). In particular, some studies have demonstrated that a stronger association between dot comparison performance and mathematics ability is found with children rather than adults (Fazio et al., 2014; Inglis et al., 2011). This has often been interpreted as a correlation between ANS acuity and mathematics achievement, but these results could also be caused by a mutual relationship with other cognitive skills, such as inhibition or visuo-spatial skills.

Second, one can conclude from this study that dot comparison tasks measure different cognitive constructs in adults in comparison to 7–11 year old children. Adults were significantly more likely than children to use numerosity information when comparing dot arrays. From this, future research

should not assume the same underlying processes contribute to accuracy scores for different developmental groups; dot comparison tasks appear to be better measures of ANS acuity for adults. However, this group was far from homogeneous, and still for 30% of adults tested, numerosity judgements did not explain significant extra variance in performance above that explained by visual characteristics.

Finally, although evidence of ANS processing was not found for all participants, this study nevertheless shows that, for some individuals, dot comparison task performance is not only based on processing the visual cues of the stimuli. In their study exploring the influence of visual characteristics on performance, Gebuis and Reynvoet (2012a) proposed that the existence of an ANS that can extract number independently from visual cues appears unlikely. The present results suggest that although this hypothesis may fit with the performance patterns of most children, there remain some children and a majority of adults who are able to process numerosity information independently from visual cue information.

The above findings were consistent across three studies including dot comparison tasks that varied in the range of numerosities represented and the number of trials completed by the participants. Additionally, for a subset of 57 adult participants, the influence of the protocol for creating the dot array stimuli was analysed. Whether the stimuli were created with either the Panamath protocol (designed to control for single visual cues, excluding convex hull), or the Gebuis and Reynvoet protocol (designed to control for multiple visual cues, including convex hull), had no influence on participants' use of numerosity information. Participants' performance was just as likely to be influenced by numerosity information independently from visual cue information on trials created with either protocol. The findings appear robust despite several methodological distinctions between tasks; nevertheless, future research should assess whether results are consistent across other versions of non-symbolic comparison tasks. Moreover, the results of this study are limited to dot comparison tasks; further research could use this method to investigate whether numerosity information is predictive of performance on other tasks designed to measure ANS acuity, such as non-symbolic estimation and non-symbolic arithmetic tasks.

8.6 Summary of findings

In sum, this study has shown that numerosity processing does not independently predict dot comparison task performance for all participants. To be precise, for the majority of children and some adults, numerosity processing did not explain significant additional variance in dot comparison task performance over and above visual cue processing. Therefore, for these participants, we do not require the hypothesis that ANS processing is involved in non-symbolic numerosity judgements to account for their behaviour. This finding has theoretical implications for research showing a correlation between non-symbolic dot comparison performance and symbolic mathematical ability, as it appears likely that this relationship may not be caused by the assumed mutual relationship with ANS acuity, especially for children.

Part V

General Discussion

Chapter 9

Conclusions

The aim of this chapter is to summarise and review the findings and conclusions from the empirical work presented in this thesis. First, a brief introduction is provided as a reminder of the current state of the ANS literature and of the overarching aims of this thesis. Following this, an overview of the main findings of each study is presented with reference to the original research questions outlined in Chapter 1. The following section provides a review of the theoretical implications of the results in relation to current ANS theory, and with regard to interpretation of conclusions gained from previous dot comparison task studies. The methodological implications of the current results are then reviewed, before a discussion of the direction of future research exploring ANS processing. Finally, a summary is provided to conclude this thesis.

9.1 Introduction

Dot comparison tasks are commonly used to measure children's and adults' ANS acuities. Many researchers have reported results obtained from these tasks with the implicit assumption that they provide valid and reliable measures of ANS acuity. Importantly, some researchers have used findings from dot comparison tasks as evidence of the ANS as a core system supporting formal mathematics (Feigenson et al., 2004). High-profile studies have reported the link between ANS acuity, as measured by a dot comparison task, and mathematics achievement (e.g. Halberda et al., 2008), and consequently researchers have attempted to develop interventions to improve mathematics

ability by training the ANS (DeWind & Brannon, 2012; Hyde et al., 2014; Park & Brannon, 2013). Although modest success has been demonstrated with non-symbolic arithmetic task training (Park & Brannon, 2013), there has not yet been any evidence to suggest that training using the dot comparison paradigm can improve mathematics achievement. This is likely due to different cognitive skills underpinning performance on the two tasks.

Despite the wide use of dot comparison tasks within the literature, very little is known about the cognitive skills that underlie task performance. Recently, studies have begun to emerge demonstrating that performance is substantially influenced by changes to the visual characteristics of the stimuli (Gebuis & Reynvoet, 2012a; Smets et al., 2014). Moreover, some researchers have hypothesised that dot comparison task performance may be entirely accounted for by visual cue processing and that there is likely to be no independent contribution from the ANS (Gebuis & Reynvoet, 2012a). Others have proposed that inhibitory control may play an important role in comparing non-symbolic dot arrays (Cappelletti et al., 2014; Fuhs & McNeil, 2013; Gilmore et al., 2013; Szűcs et al., 2015). These findings and hypotheses raise questions regarding the validity of dot comparison tasks as measures of ANS acuity, and warrant further research.

The overarching aim of this thesis was to explore the cognitive and methodological factors that influence performance on dot comparison tasks in order to establish whether they can be considered valid and reliable measures of ANS acuity.

9.2 Overview of results

A summary of the results and conclusions from each of the studies in this thesis is presented below, alongside a reminder of the corresponding original research questions.

9.2.1 Part II: Visual cues in dot comparison tasks

> **Study 1 research questions:** Do the visual cues in dot array stimuli influence task performance? Are tasks created with different controls for visual cues measuring the same cognitive construct? How reliable are these different methods?

Study 1 investigated how the accuracy and reliability of dot comparison task judgements were influenced by the visual cue control protocol used to create the stimuli. The same participants completed dot comparison task trials created with the Gebuis and Reynvoet (2011) protocol and with the Panamath protocol (Halberda et al., 2008). The results from this study showed that the visual cues in both sets of trials had a substantial influence on task performance, and significant congruency effects were found. That is, participants performed more accurately on trials where the larger numerosity had larger visual cues than trials where the larger numerosity had smaller visual cues. A novel finding from this study was that across all trials, only convex-hull congruency effects were found, and there were no significant cumulative surface area congruency effects. This result provides clarification for conflicting findings in the literature regarding overall congruency effects. Specifically, although many studies have reported significant congruency effects from dot comparison tasks (e.g. Barth et al., 2006; Cappelletti et al., 2014; Gilmore et al., 2013; Hurewitz et al., 2006; Nys & Content, 2012; Szűcs et al., 2015), several other studies have reported that they did not find such effects (e.g. Gebuis & van der Smagt, 2011; Odic, Libertus, et al., 2013; Odic et al., 2014). In line with the present findings, it is likely that the studies that previously failed to show congruency effects may have done so because they did not take into consideration the convex-hull size of the arrays in their congruency analyses.

Additionally, as part of Study 1, an analysis of the dot arrays created with the Panamath protocol demonstrated that this method of stimuli generation does not appropriately control for convex-hull size, resulting in a confound between convex-hull size and numerosity in the arrays. This means that if participants were to focus on convex-hull size alone to make their judgements, they would perform significantly above chance on trials made using this protocol.

Importantly, Study 1 found that tasks created with different controls for visual cues were not measuring the same cognitive construct. Participants' performances on the two types of trials were only weakly and non-significantly related ($r = .260$, $p = .078$). Participants' performance on the Gebuis and Reynvoet trials only explained 7% of the variance in their performance on the Panamath trials. This has important implications for the comparability of dot comparison task results generated from research groups

using different visual cue control methods to create their stimuli.

Finally, Study 1 reported that the immediate test-rest reliability differed between the two protocols, and was unacceptably low for the Panamath trials ($r = .286$). It is possible that Panamath trials are less reliable because the visual cue controls are less rigorous, and, therefore, the trials may not involve inhibition to the same extent as the Gebuis and Reynvoet (2011) trials. If inhibitory control load from the Gebuis and Reynvoet trials is higher due to the added manipulation of convex hull, it could be that the inhibition processing in the trials is the reliable element of the task.

Overall, this study resulted in several novel findings which should be considered when reviewing the findings from previously published dot comparison tasks. Panamath is a widely used tool, and many of the high-profile studies relating ANS acuity to mathematics achievement (e.g. Halberda et al., 2008; Starr et al., 2015) have employed this method of visual cue control, which appears to be unreliable which and measures different cognitive processes to tasks created using the Gebuis and Reynvoet protocol.

9.2.2 Part III: Inhibition in dot comparison tasks

> **Study 2 research questions:** How does the absolute set size, and the consequent change in the visual characteristics of dot arrays, influence non-symbolic comparison task performance? Are responses in line with an inhibitory control account of performance?

Part III of this thesis moved on to explore the potential role of inhibition in dot comparison task performance. Study 2 explored how variation in the absolute set size of dot arrays influenced participants' accuracy scores. Trials with fixed numerosity ratios were presented in four different set sizes ranging from 10 to 76 dots. The overall result was that as set size increased, participants' accuracy scores decreased. This is a novel finding in the literature which contradicts the dominant view that ANS judgements are only influenced by the ratio difference between the numerosities, and not the absolute magnitude of the values (Barth et al., 2005; Dehaene, 1997).

A second finding from this study was that visual cue congruency effects also varied with set size. Specifically, for smaller numerosity trials, participants were more influenced by the average dot size of the arrays than

convex-hull size. For larger numerosity trials, participants were more influenced by the convex-hull size of the arrays than average dot size. This makes sense when considering the way in which dot arrays are constructed. The density of the dots in an array necessarily increases with numerosity due to limited screen space, and consequently increasingly crowded dots create a more prominent boundary to the array. Therefore, convex hull becomes a more salient visual cue that is particularly difficult to inhibit in incongruent, large set size trials.

Finally, Study 2 used a regression analysis to demonstrate that both numerosity and visual cue processing contributed to participants' accuracy scores.

The combined results of this study are in line with the inhibition-based competing processes hypothesis. Inhibition is likely to be involved in the processing of dot comparison task trials where visual cues are incongruent with the numerosity represented in the array. More salient visual cues caused by changes in set size were found to lead to a higher inhibition load as measured by congruency effects. Finally, the finding that overall accuracy scores varied with set size is an important result that should be taken into account when designing or comparing dot comparison task methodologies.

> **Study 3 research question:** Does dot comparison task performance follow the same pattern of results as classic inhibition tasks?

The aim of Study 3 was to provide further evidence for the role of inhibition in dot comparison task judgements. Previous results from the inhibition literature show greater interference on the initial one or two incongruent Stroop task trials than subsequent incongruent trials (Henik et al., 1997). In order to demonstrate that individuals' responses to dot comparison trials follow a pattern of response similar to that of classic inhibition tasks, the frequency of conflict between congruent and incongruent trials was manipulated. Incongruent 'test trials' were preceded either by blocks of similarly incongruent trials, or by blocks of contrasting congruent trials. As expected, results showed that performance on the incongruent test trials was significantly higher when there was no conflict in congruency status from the preceding block of trials.

Study 3, therefore, provided further evidence in support of the hypothesis that inhibition is involved in dot comparison task processing.

> **Study 4 and Study 5 research question:** Does dot comparison task performance correlate with inhibition task performance?

The aim of both Studies 4 and 5 was to provide evidence of a direct link between individual differences in dot comparison task performance and an explicit measure of inhibitory control. Study 4 investigated dot comparison task performance alongside a colour-word Stroop task considered to be a standard measure of interference control inhibition. Although both tasks generated substantial congruency effects, these congruency effects were found not to be significantly correlated across tasks. This was a puzzling result in light of previous findings reported in this thesis. However, research by Shilling et al. (2002) has shown that the lack of a significant correlation across inhibition tasks could be due to differences in the surface characteristics of the measures. Specifically, Shilling et al. (2002) found that the stimulus and response dimensions of tasks designed to measure the same cognitive construct (i.e. interference control), must be highly similar to produce a relationship between individual differences across tasks.

In response to this, Study 5 investigated dot comparison task performance alongside three different measures of interference control that were more similar in terms of both stimulus and response formats. The inhibition tasks included an animal size Stroop, a number size Stroop, and a classic Flanker task. Crucially, this study found that individual differences in congruency effects obtained on the dot comparison task significantly correlated with congruency effects on the animal and number size Stroop tasks, but did not correlate with congruency effects on the Flanker task. This result is in line with Shilling et al.'s (2002) finding, as the two Stroop variations were more similar in format to the dot comparison task. These tasks involved the comparison of two visually distinct images where the stimuli's physical sizes were unhelpful for the demands of the task. In contrast, the Flanker task involved the processing of a single image in the centre of the screen, and the inhibition of surrounding stimuli with unhelpful semantic value, rather than physical size.

Nevertheless, Shilling et al.'s result that such small variations in task formats can influence the correlation between inhibition measures is consis-

tent with Study 5's finding that dot comparison task performance was found to be significantly related to two separate measures of interference control. Combined with the results of Study 2 and Study 3, the findings of Study 5, and ultimately Part III of this thesis, provide substantial evidence that success on dot comparison tasks requires inhibitory control skills.

9.2.3 Part IV: Do non-symbolic numerosity tasks involve numerosity processing?

Re-analysis of data research question: Do dot comparison tasks involve numerosity processing at all?

The final study presented in this thesis investigated whether numerosity plays a role in dot comparison judgements over and above visual cue processing. Given the evidence provided in Parts II and Part III, it is clear that visual cue processing has a substantial influence on dot comparison task judgements. Moreover, previous studies by Gebuis and colleagues have suggested that visual cue processing may entirely account for individual differences in dot comparison performance (Gebuis & Reynvoet, 2012a, 2012b). The regression analysis provided as part of Study 2 demonstrated initial evidence to suggest that both visual cues *and* numerosity processing influenced task accuracy. However, the analysis was not hierarchical and therefore could not provide any insight on whether numerosity information was processed *additionally* to visual cues.

The re-analysis presented in Chapter 8 used a hierarchical regression to demonstrate that there are individual differences in the use of numerosity information, over and above visual cues, in dot comparison tasks. Specifically, developmental differences were found demonstrating that for almost 70% of children and 30% of adults, the numerosity information in the trials did not explain significant additional variance in their accuracy scores over and above that explained by the visual cues. This finding was robust across three different tasks with methodological differences including variation in the numbers of trials, numerosity ranges, ratios and visual cue controls.

In sum, the re-analysis study presented in Part IV of this thesis reported novel evidence to show that numerosity processing is not independently involved in dot comparison judgements for all participants. This finding has critical implications for the use of dot comparison tasks as a measure of ANS

acuity, particularly for use with children.

9.3 Theoretical implications

9.3.1 Implications for ANS theory

The results of this thesis undoubtedly have implications for current ANS theory. As first described in the literature review (Chapter, 1, Section 1.2), the original and dominant model of the ANS proposes that numerical representations are abstract by nature and formed independently of non-numerical factors (Feigenson et al., 2004). The signature of the ANS is ratio-dependent performance, with accuracy on comparison tasks decreasing as the ratio between to to-be-compared numerosities approaches 1 (Barth et al., 2005; Dehaene, 1997). According to this account, the only influences on approximate numerosity judgements are ratio effects and the acuity of the individual's ANS representations.

If dot comparison task performance is assumed to be a pure measure of ANS acuity, the studies in the current thesis conflict with this theory in at least three ways. First, Study 1 showed that ANS representations are not independent of non-numerical factors, but are significantly influenced by the visual characteristics of the stimuli. Second, Study 2 showed that ratio differences and ANS acuity are not the only influence on dot comparison task performance. The results of Study 2 demonstrated that variation in the absolute magnitude of dot arrays influenced task accuracy, whilst numerosity ratios were kept constant. Third, Study 3 highlighted the significant influence of the congruency status of preceding trials on subsequent judgements of numerosity. All of these findings are at odds with the premise that dot comparison tasks provide a valid measure of ANS acuity as described by the standard model.

Prior to the work in this thesis, other researchers had raised issues with the standard account of dot comparison task performance, and suggested that performance may be explained without reference to the ANS. Gebuis and colleagues proposed that numerosity judgements on non-symbolic comparison tasks may be made solely by weighing up multiple visual cues in the stimuli (Gebuis & Reynvoet, 2012a, 2012b). The authors propose that given the strong relation between number and visual cues in real life, it is unlikely that an ANS exists that is independent of these non-numerical cues.

Gebuis and Reynvoet (2012a) give the example of a bag with more apples that looks physically larger than a bag with fewer apples in it. They suggest that visual cues and number are nearly always confounded in everyday life, and question why an evolutionary, innate system would require approximate quantity processing to be independent from visual cues. The results of the present thesis strongly support Gebuis and Reynvoet's view that multiple visual cues are taken into account whilst making quantity judgements, but the results of the re-analysis study in Chapter 8 demonstrate the additional role of numerosity processing for some participants. Therefore, Gebuis and Reynvoet's theory may stand for some participants, particularly children who are less likely to use numerical cues over and above visual cues. However, for many adults this theory appears insufficient. The re-analysis study reported here showed that 70% of adults' accuracy scores were significantly influenced by numerosity information over and above visual cues.

More recently, an inhibition-based account of dot comparison performance has been proposed by several ANS researchers (Cappelletti et al., 2014; Fuhs & McNeil, 2013; Gilmore et al., 2013; Nys & Content, 2012; Szűcs et al., 2015). This account suggests that dot comparison judgements may involve a mixture of ANS processing and inhibitory control. Individuals may attempt to judge which array is more numerous (ANS processing), but visual cues may compete with this initial judgement (competing processes), and inhibition may be required to inhibit a response based on misleading visual cues. The results of the present thesis align with this proposal that inhibition is involved in the processing of incongruent dot comparison task trials. All studies showed a significant congruency effect, with higher performance on trials where the visual cues were congruent with numerosity, in comparison to trials where the visual cues were incongruent with numerosity. Congruency effects are a key signature of all inhibition tasks, and Study 3 additionally demonstrated how dot comparison responses followed the same pattern of results as classic inhibition tasks when the frequency of conflict between congruent and incongruent trials was manipulated. Finally, Study 5 provided further evidence for the role of inhibition, demonstrating a correlation between participants' dot comparison congruency effects and congruency effects measured on two different interference control inhibition tasks. Interestingly, due to the heavy inhibition load of incongruent trials, a dot comparison task has recently been used as a measure of inhibitory

control skills in a study investigating the role of inhibition in different components of arithmetic (Gilmore, Keeble, Richardson, & Cragg, 2015).

A very recent study by DeWind et al. (2015) highlighted the flaws of previous ANS models that do not account for non-numerical cue processing, and proposed a new model intended to account for the contribution of visual features. DeWind et al. (2015) propose that their new model captures variance in dot comparison task behaviours that were previously unaccounted for, therefore providing a valid and reliable estimate of w that remains constant across congruent and incongruent trials. Although this model provides support for the importance of visual cue processing in dot comparison tasks, the authors conclude that numerosity processing is a more influential factor on performance than visual cue processing. Due to the different analysis techniques, it is difficult to assess whether this finding is in conflict with the results of the re-analysis study presented in Chapter 8 of this thesis. DeWind et al.'s conclusions were based on the fact that performance could not be better explained by the discrimination of a single non-numerical feature of the stimuli, whereas the re-analysis of data presented here assessed the relative influence of visual cues as a whole. Nevertheless, this model may provide a tool for identifying the different effects of numerical and visual cue features of dot comparison task stimuli in the future (DeWind et al., 2015).

9.3.2 Implications for the relationship between ANS acuity and mathematics achievement

The overall results of this thesis have implications for proposal that the ANS is a core system that supports formal mathematics skills (Feigenson et al., 2004). Several studies have suggested that there is a causal relationship between ANS acuity and formal mathematics achievement (Libertus et al., 2011; Mazzocco et al., 2011b; Piazza et al., 2010, 2013). However, these studies have assumed that non-symbolic comparison tasks provide a valid and reliable measure of ANS acuity. The results of Study 1 suggest that Panamath, a method of stimuli generation that does not control for convex-hull size, may not create dot comparison trials that provide adequate task reliability. Many of the studies that have investigated the relationship between ANS acuity and mathematics achievement have used Panamath to generate their stimuli (e.g. Fazio et al., 2014; Halberda et al., 2008, 2012; Libertus et al., 2013a), thus the validity of such conclusions can be questioned. Indeed,

most studies exploring the relationship between dot comparison task performance and mathematics achievement use a method of visual cue control that does not explicitly control for convex-hull size. These studies are therefore also likely to contain a convex-hull confound with numerosity whereby the larger numerosities in the trials also have larger convex hulls. This confound means that participants may be able to perform above chance level by focusing on the visual cues of the arrays alone, and consequently it would not be possible to decipher whether a relationship with mathematics achievement was due to a mutual correlation with ANS acuity or the ability to make visual cue judgements. Furthermore, the reliability of a dot comparison task that does not control for convex-hull size may also be suboptimal, as was found with the Panamath trials in Study 1. Table 9.1 below presents summaries of the studies that have explored the correlation between dot comparison task performance and formal mathematics ability (as presented initially in Section 1.4 of the literature review, Chapter 1), with the studies that have used dot comparison stimuli created without systematic convex-hull size controls highlighted in grey. It could be argued that the results of the highlighted studies are questionable given the evidence summarised above demonstrating the importance of controlling convex-hull size for dot comparison task reliability and validity.

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Agrillo et al. (2013)	Sequential	Adults	Acc	Mental arithmetic	$r = .463^{**}$
	Sequential	Adults	Acc	Mathematical Reasoning (WAIS-R)	$r = .489^{**}$
	Sequential	Adults	RT	Mental arithmetic	$r = .391^{*}$
	Sequential	Adults	RT	Mathematical Reasoning (WAIS-R)	$r = .449^{**}$
Bartelet et al. (2014)	Simultaneous	Children	RT	Arithmetic fact retrieval (TTA)	$r = -.14$
	Simultaneous	Children	Acc	Arithmetic fact retrieval (TTA)	$r = .24^{*}$
Bonny and Lourenco (2013)	Simultaneous	Children	ANS precision (predicted for untested ratio)	TEMA-3	$r = .387^{***}$

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Brankaer et al. (2014)	Simultaneous	Children (6 years)	Acc	Tempo Test Arithmetic	$r = .36^{*}$
	Simultaneous	Children (6 years)	Acc	Curriculum-based standardised test	$r = .15$
	Simultaneous	Children (6 years)	RT	Tempo Test Arithmetic (TTA)	$r = -.13$
	Simultaneous	Children (6 years)	RT	Curriculum-based standardised test	$r = .02$
	Simultaneous	Children (8 years)	Acc	Tempo Test Arithmetic (TTA)	$r = .14$
	Simultaneous	Children (8 years)	Acc	Curriculum-based standardised test	$r = -.16$
	Simultaneous	Children (8 years)	RT	Tempo Test Arithmetic (TTA)	$r = -.17$
	Simultaneous	Children (8 years)	RT	Curriculum-based standardised test	$r = -.20$

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Fazio et al. (2014)	Simultaneous	Children	w and RT combined	School mathematics assessment (PSSA) score	$r = .60^{*}$
Fuhs and McNeil (2013)	Simultaneous	Children	Acc	TEMA-3	$r = .19$
Gilmore et al. (2013)	Simultaneous	Children	Acc	WJ-III Calculation subtest	$r = .57^{***}$
Guillaume et al. (2013)	Simultaneous	Adults	w	Addition arithmetic RT	$r = .47^{**}$
Halberda et al. (2008)[1]	Intermixed	Children (5 years)	w	TEMA-2	$r = .370^{**}$

[1]Dot comparison performance measured at 14 years, mathematics achievement measured at different time points provided in table.

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
	Intermixed	Children (5 years)	w	WJ-Rcalc	$r = .356^{**}$
	Intermixed	Children (6 years)	w	TEMA-2	$r = .374^{**}$
	Intermixed	Children (6 years)	w	WJ-Rcalc	$r = .571^{***}$
	Intermixed	Children (7 years)	w	TEMA-2	$r = .488^{***}$
	Intermixed	Children (8 years)	w	TEMA-2	$r = .569^{***}$
	Intermixed	Children (8 years)	w	WJ-Rcalc	$r = .531^{***}$
	Intermixed	Children (9 years)	w	WJ-Rcalc	$r = .498^{***}$
	Intermixed	Children (10 years)	w	WJ-Rcalc	$r = .342^{**}$
	Intermixed	Children (11 years)	w	WJ-Rcalc	$r = .501^{***}$

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Halberda et al. (2012)	Intermixed	Children, Adults	w	Self-reported school mathematics achievement	$r = -.19^{***}$
	Intermixed	Children, Adults	RT	Self-reported school mathematics achievement	$r = -.09^{***}$
Holloway and Ansari (2009)	Simultaneous	Children	NDE	WJ-III Mathematics Fluency and Calculation composite	$r = -.015$
Inglis et al. (2011)	Simultaneous	Children	w	WJ-III Calculation subtest	$r = -.548^{**}$[2]
	Simultaneous	Adults	w	WJ-III Calculation subtest	$r = .161$[2]

[2]Partial correlation controlling for non-verbal IQ and age.

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Kolkman et al. (2013)	Simultaneous	Children	Acc	Standardised mathematics test	$r = .16$
Libertus et al. (2011)	Simultaneous	Children	Acc	TEMA-3	$r = -.424^{***}$
	Simultaneous	Children	w	TEMA-3	$r = -.265^{**}$
	Simultaneous	Children	RT	TEMA-3	$r = -.283^{***}$
Libertus et al. (2012)	Simultaneous	Adults	w	Scholastic Aptitude Test (SAT) Quantitative	$r = -.22^{*}$
Libertus et al. (2013a)	Simultaneous	Children	Acc	TEMA-3	$r = .52^{**}$
	Simultaneous	Children	w	TEMA-3	$r = -.42^{**}$
	Simultaneous	Children	RT	TEMA-3	$r = -.36^{**}$

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Libertus et al. (2013b)	Simultaneous	Children	Acc	TEMA-3 informal mathematics items	$r = .44^{***}$
	Simultaneous	Children	Acc	TEMA-3 formal mathematics items	$r = .06$
Lonnemann et al. (2015)	Simultaneous	Children	NDE	Addition arithmetic	$r = -.04$
	Simultaneous	Children	NDE	Subtraction arithmetic	$r = .01$
Lourenco et al. (2012)	Intermixed	Adults	Acc	WJ-III Calculation subtest	$r = .320^{**}$
	Intermixed	Adults	Acc	KeyMath 3 Geometry subtest	$r = .332^{***}$
Lyons et al. (2014)	Simultaneous	Children	Acc and RT combined	Tempo Test Automatiseren (TTA)	$r = .554^{***}$

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Lyons and Beilock (2011)	Simultaneous	Adults	w	Mental arithmetic	$r = -.339^{*}$
Mazzocco et al. (2011b)[3]	Simultaneous	Children	Acc	TEMA-3	$r = -.527^{*}$
	Simultaneous	Children	w	TEMA-3	$r = -.456$
Mundy and Gilmore (2009)	Simultaneous	Children	Acc	Curriculum-based mathematics test	$r = .35$
	Simultaneous	Children	NDE	Curriculum-based mathematics test	$r = .02$
Nys and Content (2012)	Simultaneous	Adults	Acc	Tempo Test Rekenen (TTR)	$r = .16$

[3]Dot comparison performance measured at age 3–4 years (scores adjusted for age and display time at initial testing), TEMA-3 measured at 6-7 years (scores adjusted for age and grade at follow-up testing).

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
	Simultaneous	Adults	RT	Tempo Test Rekenen (TTR)	$r = -.08$
Price et al. (2012)	Sequential	Adults	NRE	WJ Math Fluency subtest	$r = .01$
	Simultaneous	Adults	NRE	WJ Math Fluency subtest	$r = .01$
	Intermixed	Adults	NRE	WJ Math Fluency subtest	$r = .03$
	Sequential	Adults	w	WJ Math Fluency subtest	$r = .10$
	Simultaneous	Adults	w	WJ Math Fluency subtest	$r = -.28$
	Intermixed	Adults	w	WJ Math Fluency subtest	$r = -.24$
Sasanguie et al. (2011)	Simultaneous	Children	RT/Error	Curriculum-based standardised test	$r = -.16$[4]
	Simultaneous	Children	NDE	Curriculum-based standardised test	$r = .08$[4]

[4]Partial correlation controlling for grade (year group).

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Sasanguie et al. (2012)	Simultaneous	Children	RT/Error	Curriculum-based standardised test	$r = -.18$[4]
	Simultaneous	Children	NDE	Curriculum-based standardised test	$r = -.12$[4]
Sasanguie et al. (2013)	Simultaneous	Children	Acc	Tempo Test Rekenen (TTR)	$r = .14$[5]
	Simultaneous	Children	w	Tempo Test Rekenen (TTR)	$r = -.17$[5]
	Simultaneous	Children	Acc	Curriculum-based standardised test	$r = .09$[5]
	Simultaneous	Children	w	Curriculum-based standardised test	$r = -.17$[5]

[5]Partial correlation controlling for grade (year group) and spelling achievement.

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
Soto-Calvo et al. (2015)	Simultaneous	Children	Acc	WIAT-II Mathematical Reasoning subtest	$r = .34^{***}$
	Simultaneous	Children	Acc	WIAT-II Numerical Operations subtest	$r = .39^{***}$
Starr et al. (2015)	Simultaneous	Children	w	TEMA-3	$r = -.42^{**}$
Vanbinst et al. (2012)	Simultaneous	Children	NDE	Curriculum-based standardised test	$r = .03$
Zhou et al. (2015)	Simultaneous	Children (8 years)	acc	School achievement test	$r = .28^{**}$
	Simultaneous	Children (8 years)	RT	School achievement test	$r = .24^{**}$
	Simultaneous	Children (9 years)	acc	School achievement test	$r = .18^{*}$

Study	Stimuli presentation	Age group	Index	Math measure	Correlation
	Simultaneous	Children (9 years)	RT	School achievement test	$r = .03$
	Simultaneous	Children (10 years)	acc	School achievement test	$r = .25^{**}$
	Simultaneous	Children (10 years)	RT	School achievement test	$r = .06$

Table 9.1: A summary of the studies that have reported the relationship between non-symbolic comparison task performance and formal mathematics abilities in a typical population (both adults and children). The Pearson's correlation coefficients are provided, along with key characteristics of the studies including the stimuli presentation method, the age group of the participants, the index of non-symbolic comparison performance employed, and the mathematics ability measure. The studies that did not systematically control for convex-hull size are highlighted in grey. Acc = accuracy, RT = response time, w = Weber fraction, NDE = numerical distance effect, NRE= numerical ratio effect. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

The findings reported in this thesis also show that inhibition plays an important role in dot comparison task performance. This evidence supports the view that inhibition could be a key mediator in the link between dot comparison task performance and mathematics achievement (Fuhs & McNeil, 2013; Gilmore et al., 2013). If this is the case, then the development of successful interventions to improve participants' formal mathematical skills through dot comparison training may be unachievable. If inhibition skills play a critical role in non-symbolic comparison processing and also mathematics performance, then it could be argued that practice on these tasks may still lead to an improvement in dot comparison performance and, in turn, mathematics performance by improving inhibition skills. However, there is scarce evidence in the literature to suggest that training on inhibition tasks can successfully lead to improvements and subsequent transfer effects to non-trained tasks (Enge et al., 2014; Thorell, Lindqvist, Bergman Nutley, Bohlin, & Klingberg, 2009), and so this hypothesis remains unlikely to be true.

Indeed, to date, only non-symbolic arithmetic task training has demonstrated any transfer to measures of mathematical ability (Park & Brannon, 2013). It is possible, in this case, that the transfer effect stemmed from the arithmetic element of the task demands and that different cognitive processing is required to add and subtract approximate quantities than to simply compare quantities.

A study by Hyde and colleagues (2014) has shown that practice on dot comparison task trials leads to a significant improvement in children's response times when completing arithmetic problems, but no improvement in accuracy scores. In fact, the dot comparison task training in this study failed to lead to improvements in a subsequent dot comparison task itself, and so it is likely that the response time decrease in the arithmetic task was due to factors other than ANS acuity, such as familiarity with the study procedures. Nevertheless, Hyde et al. (2014) only included 60 trials in their training program, and so it is possible that the effectiveness of the training was limited by this short training exposure. It remains to be seen whether a training study with a more substantial period of practice will lead to formal mathematical gains.

In sum, the empirical results provided in this thesis cast doubt over the validity of previous findings relating to the correlation between non-symbolic

comparison task performance and mathematics achievement. Many of the significant positive correlations that have been reported in the literature to date could potentially be caused by mutual correlations with other cognitive skills, e.g. inhibition or visual processing, rather than ANS acuity. Indeed, very few studies have controlled for inhibition as a potential mediating variable. Furthermore, many of the studies in the field are methodologically flawed, given that they fail to adequately control for a highly influential visual cue, convex-hull size. Future research will need to give consideration to these two factors in order to effectively evaluate the relationship between ANS acuity and mathematics achievement using the dot comparison task.

9.4 Methodological implications

The results presented in this thesis show that multiple methodological variables influence participants' dot comparison task performance. The factors explored in this thesis include the absolute magnitude of the dot arrays, the frequency of conflict between congruent and incongruent trials and, importantly, the way in which visual cues are controlled in the stimuli. Considering the substantial influence of these factors on accuracy scores, a standardised methodology may be beneficial for future research. Dietrich et al. (2015) have highlighted many more methodological inconsistencies within the dot comparison task literature and provided a checklist for developing new tasks. However, the influence of some of these factors has either not yet been systematically explored, or the recommendations remain vague. For example, Dietrich et al. recommended the use of numerosities over the subitising range to create stimuli (i.e. 4+), but did not provide any detailed or specific advice further to this.

At a minimum, the results of Study 1 demonstrate that the visual cues, including both dot size and convex-hull size, should be controlled for when generating dot comparison stimuli. Researchers should use the concrete values of the visual characteristics in their stimuli to perform post-hoc analyses to ensure there are no confounds with numerosity present (Gebuis & Reynvoet, 2011). Researchers may also benefit from using these values in their data analyses to gain a greater insight into individual differences in performance. In particular, if researchers are to continue to use dot comparison tasks as a measure of ANS acuity, then an analysis similar to the re-analysis

of data described in Chapter 8 may be required to disentangle the contribution of independent numerosity processing from visual cue processing.

Nevertheless, considering the emergence of new research highlighting substantial issues concerning the reliability and validity of different dot comparison task methodologies (Study 1; Inglis & Gilmore, 2014; Smets et al., 2015), and given the relatively small influence of numerosity processing over and above visual cues, it is difficult to see the benefits of continuing to use dot comparison tasks as a measure of ANS acuity.

9.5 Future research

In order to advance our understanding of the ANS and its correlates, future research may benefit from a shift in focus towards the development of alternative protocols to measure ANS acuity. Due to the unwanted influence of visual cues, the use of dot comparison tasks may not be appropriate. Cross-modal methods involving a mixture of visual dot arrays and auditory stimuli have successfully been used in previous research. Barth et al. (2005) found that children were able to integrate quantity information from these two different modalities, demonstrating that performance in a dual-modality task was not significantly different to performance in a single visual modality task. It is possible this method of non-symbolic comparison may require less inhibitory control demands than standard dot comparison tasks, and therefore provide a more valid measure of ANS acuity.

A review published by De Smedt et al. (2013) noted that the relationship between symbolic, rather than non-symbolic, numerical magnitude processing and mathematics achievement appears to be robust. Studies investigating individual differences in approximate judgements in a symbolic format do not appear to be subject to the same constraints as non-symbolic tasks (De Smedt et al., 2013). In terms of developing interventions to improve formal mathematics achievement, research exploring symbolic magnitude processing is likely to be more successful.

9.6 Summary

To conclude, the studies reported in this thesis have provided novel evidence to show that dot comparison task judgements are substantially influenced

by multiple methodological variables. Following this, results obtained from dot comparison tasks created with diverse procedures do not appear to be measuring the same underlying cognitive processes, as is implicitly assumed in the literature. Moreover, a key finding from this thesis revealed that dot comparison tasks do not measure numerical processing skills independently from visual cues for the majority of children, and some adults. Together, these findings raise doubt over the future use of dot comparison tasks as measures valid and reliable measures of the Approximate Number System.

References

Agrillo, C., Piffer, L., & Adriano, A. (2013). Individual differences in non-symbolic numerical abilities predict mathematical achievements but contradict ATOM. *Behavioral and Brain Functions*, *9*(1), 26.

Andersson, U. (2010). Working memory as a predictor of written arithmetical skills in children: The importance of central executive functions. *British Journal of Educational Psychology*, *78*(2), 181–203.

Ansari, D., Lyons, I. M., van Eimeren, L., & Xu, F. (2007). Linking visual attention and number processing in the brain: The role of the temporo-parietal junction in small and large symbolic and nonsymbolic number comparison. *Journal of Cognitive Neuroscience*, *19*(11), 1845–1853.

Ashcraft, M. H., & Battaglia, J. (1978). Cognitive arithmetic: Evidence for retrieval and decision processes in mental addition. *Journal of Experimental Psychology: Human Learning & Memory*, *4*(5), 527–538.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556–559.

Banks, J., & Oldfield, Z. (2007). Understanding Pensions: Cognitive Function, Numerical Ability and Retirement Saving. *Fiscal Studies*, *28*(2), 143–170.

Barbaresi, W. J., Katusic, S. K., Colligan, R. C., Weaver, A. L., & Jacobsen, S. J. (2005). Math learning disorder: Incidence in a population-based birth cohort, 1976–82, Rochester, Minn. *Ambulatory Pediatrics*, *5*(5), 281–289.

Baroody, A. J. (1994). An evaluation of evidence supporting fact-retrieval models. *Learning and Individual Differences*, *6*(1), 1–36.

Baroody, A. J. (2003). The development of adaptive expertise and flexibility: The integration of conceptual and procedural knowledge. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts*

and skills: Constructing adaptive expertise (pp. 1–34). Mahwah, NJ: Erlbaum.

Bartelet, D., Vaessen, A., Blomert, L., & Ansari, D. (2014). What basic number processing measures in kindergarten explain unique variability in first-grade arithmetic proficiency? *Journal of Experimental Child Psychology*, *117*, 12–28.

Barth, H., Beckmann, L., & Spelke, E. S. (2008). Nonsymbolic, approximate arithmetic in children: Abstract addition prior to instruction. *Developmental Psychology*, *44*(5), 1466–1477.

Barth, H., Kanwisher, N., & Spelke, E. S. (2002). The construction of large number representations in adults. *Cognition*, *86*, 201–221.

Barth, H., La Mont, K., Lipton, J., Dehaene, S., Kanwisher, N., & Spelke, E. (2006). Non-symbolic arithmetic in adults and young children. *Cognition*, *98*(3), 199–222.

Barth, H., La Mont, K., Lipton, J., & Spelke, E. S. (2005). Abstract number and arithmetic in preschool children. *Proceedings of the National Academy of Sciences*, *102*(39), 14116–14121.

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, *78*(2), 647–663.

Bonny, J. W., & Lourenco, S. F. (2013). The approximate number system and its relation to early math achievement: Evidence from the preschool years. *Journal of Experimental Child Psychology*, *114*(3), 375–388.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624.

Brankaer, C., Ghesquière, P., & De Smedt, B. (2014). Children's mapping between non-symbolic and symbolic numerical magnitudes and its association with timed and untimed tests of mathematics achievement. *PLoS ONE*, *9*(4), e93565.

Braver, T. S. (2012). The variable nature of cognitive control: a dual mechanisms framework. *Trends in Cognitive Sciences*, *16*(2), 105–112.

Bull, R., & Espy, K. A. (2006). Working memory, executive functioning, and children's mathematics. In S. J. Pickering (Ed.), *Working memory*

and education (pp. 94–123). Burlington, MA: Academic Press.

Bull, R., Espy, K. A., & Wiebe, S. A. (2008). Short-term memory, working memory, and executive functioning in preschoolers: Longitudinal predictors of mathematical achievement at age 7 years. *Developmental Neuropsychology*, *33*(3), 205–228.

Bull, R., & Johnston, R. S. (1997). Children's arithmetical Difficulties: Contributions from processing Speed, item identification, and short-term memory. *Journal of Experimental Child Psychology*, *65*(1), 1–24.

Bull, R., Johnston, R. S., & Roy, J. A. (1999). Exploring the roles of the visualspatial sketch pad and central executive in children's arithmetical skills: Views from cognition and developmental neuropsychology. *Developmental Neuropsychology*, *15*(3), 421–442.

Bull, R., & Lee, K. (2014). Executive functioning and mathematics achievement. *Child Development Perspectives*, *8*(1), 36–41.

Bull, R., & Scerif, G. (2001). Executive functioning as a predictor of children's mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology*, *19*(3), 273–293.

Cappelletti, M., Didino, D., Stoianov, I., & Zorzi, M. (2014). Number skills are maintained in healthy ageing. *Cognitive Psychology*, *69*, 25–45.

Castronovo, J., & Göbel, S. M. (2012). Impact of high mathematics education on the number sense. *PLoS ONE*, *7*(4), e33832.

Chen, Q., & Li, J. (2014). Association between individual differences in non-symbolic number acuity and math performance: A meta-analysis. *Acta Psychologica*, *148*, 163–172.

Clayton, S., & Gilmore, C. (2014). Inhibition in dot comparison tasks. *ZDM*, *47*(5), 759–770.

Clayton, S., Gilmore, C., & Inglis, M. (2015). Dot comparison stimuli are not all alike: The effect of different visual controls on ANS measurement. *Acta Psychologica*, *161*, 177–184.

Cragg, L., & Gilmore, C. (2014). Skills underlying mathematics: The role of executive function in the development of mathematics proficiency. *Trends in Neuroscience and Education*, *3*(2), 63–68.

Crocker, L. M., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Rinehard and Winston.

Defever, E., Reynvoet, B., & Gebuis, T. (2013). Task-and age-dependent effects of visual stimulus properties on children's explicit numerosity

judgments. *Journal of Experimental Child Psychology*, *116*(2), 216–233.

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. Oxford: Oxford University Press.

Dehaene, S., Izard, V., & Piazza, M. (2005). Control over non-numerical parameters in numerosity experiments. *Unpublished manuscript (available on www.unicog.org)*.

Dempster, F. N. (1992). The rise and fall of the inhibitory mechanism: Toward a unified theory of cognitive development and aging. *Developmental Review*, *12*(1), 45–75.

De Smedt, B., Noël, M.-P., Gilmore, C., & Ansari, D. (2013). How do symbolic and non-symbolic numerical magnitude processing skills relate to individual differences in children's mathematical skills? A review of evidence from brain and behavior. *Trends in Neuroscience and Education*, *2*(2), 48–55.

Desoete, A., Ceulemans, A., Roeyers, H., & Huylebroeck, A. (2009). Subitizing or counting as possible screening variables for learning disabilities in mathematics education or learning? *Educational Research Review*, *4*(1), 55–66.

DeStefano, D., & LeFevre, J. A. (2004). The role of working memory in mental arithmetic. *European Journal of Cognitive Psychology*, *16*(3), 353–386.

DeWind, N. K., Adams, G. K., Platt, M. L., & Brannon, E. M. (2015). Modeling the approximate number system to quantify the contribution of visual stimulus features. *Cognition*, *142*, 247–265.

DeWind, N. K., & Brannon, E. M. (2012). Malleability of the approximate number system: effects of feedback and training. *Frontiers in Human Neuroscience*, *6*(68), 1–10.

Dietrich, J. F., Huber, S., & Nuerk, H.-C. (2015). Methodological aspects to be considered when measuring the approximate number system (ANS) – a research review. *Frontiers in Psychology*, *6*(295).

Emmerton, J. (1998). Numerosity differences and effects of stimulus density on pigeons' discrimination performance. *Animal Learning & Behavior*, *26*(3), 243–256.

Enge, S., Behnke, A., Fleischhauer, M., Küttler, L., Kliegel, M., & Strobel, A. (2014). No evidence for true training and transfer effects after

inhibitory control training in young healthy adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(4), 987.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*(1), 143–149.

Espy, K. A., McDiarmid, M. M., Cwik, M. F., Stalets, M. M., Hamby, A., & Senn, T. E. (2004). The contribution of executive functions to emergent mathematic skills in preschool children. *Developmental Neuropsychology*, *26*(1), 465–486.

Fazio, L. K., Bailey, D. H., Thompson, C. A., & Siegler, R. S. (2014). Relations of different types of numerical magnitude representations to each other and to mathematics achievement. *Journal of Experimental Child Psychology*, *123*, 53–72.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314.

Feigenson, L., Libertus, M. E., & Halberda, J. (2013). Links between the intuitive sense of number and formal mathematics ability. *Child Development Perspectives*, *7*(2), 74–79.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of Cross-Battery Assessment*. New York: Wiley.

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General*, *133*(1), 101–135.

Friso-van den Bos, I., van der Ven, S. H. G., Kroesbergen, E. H., & van Luit, J. E. H. (2013). Working memory and mathematics in primary school children: A meta-analysis. *Educational Research Review*, *10*, 29–44.

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Hamlett, C. L., Seethaler, P. M., ... Schatschneider, C. (2010). Do different types of school mathematics development depend on different constellations of numerical versus general cognitive abilities? *Developmental Psychology*, *46*(6), 1731–1746.

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Schatschneider, C., Hamlett, C. L., ... Changas, P. (2013). Effects of first-grade number knowledge tutoring with contrasting forms of practice. *Journal of Educational Psychology*, *105*(1), 58–77.

Fuhs, M. W., & McNeil, N. M. (2013). ANS acuity and mathematics ability in preschoolers from lowincome homes: Contributions of inhibitory control. *Developmental Science*, *16*(1), 136–148.

Garavan, H., Ross, T. J., & Stein, E. A. (1999). Right hemispheric dominance of inhibitory control: An event-related functional MRI study. *Proceedings of the National Academy of Sciences*, *96*(14), 8301–8306.

Gathercole, S. E., Pickering, S. J., Knight, C., & Stegmann, Z. (2004). Working memory skills and educational attainment: Evidence from national curriculum assessments at 7 and 14 years of age. *Applied Cognitive Psychology*, *18*(1), 1–16.

Geary, D. C., Hamson, C. O., & Hoard, M. K. (2000). Numerical and arithmetical cognition: A longitudinal study of process and concept deficits in children with learning disability. *Journal of Experimental Child Psychology*, *77*(3), 236–263.

Geary, D. C., Hoard, M. K., & Nugent, L. (2012). Independent contributions of the central executive, intelligence, and in-class attentive behavior to developmental change in the strategies used to solve addition problems. *Journal of Experimental Child Psychology*, *113*(1), 49–65.

Gebuis, T., & Gevers, W. (2011). Numerosities and space; indeed a cognitive illusion! A reply to de Hevia and Spelke (2009). *Cognition*, *121*(2), 248–252.

Gebuis, T., Kadosh, R. C., de Haan, E., & Henik, A. (2009). Automatic quantity processing in 5-year olds and adults. *Cognitive Processing*, *10*(2), 133–142.

Gebuis, T., & Reynvoet, B. (2011). Generating nonsymbolic number stimuli. *Behavior research methods*, *43*(4), 981–986.

Gebuis, T., & Reynvoet, B. (2012a). The interplay between nonsymbolic number and its continuous visual properties. *Journal of Experimental Psychology: General*, *141*(4), 642–648.

Gebuis, T., & Reynvoet, B. (2012b). The role of visual information in numerosity estimation. *PLoS ONE*, *7*(5), e37426.

Gebuis, T., & van der Smagt, M. J. (2011). False approximations of the approximate number system? *PLoS ONE*, *6*(10), e25405.

Gelman, R., & Gallistel, C. R. (1978). *The Child's Understanding of Number*. Cambridge, MA: Harvard University Press.

Gerardi, K., Goette, L., & Meier, S. (2013). Numerical ability predicts

mortgage default. *Proceedings of the National Academy of Sciences*, *110*(28), 11267–11271.

Gilmore, C., Attridge, N., Clayton, S., Cragg, L., Johnson, S., Marlow, N., ... Inglis, M. (2013). Individual differences in inhibitory control, not non-verbal number acuity, correlate with mathematics achievement. *PLoS ONE*, *8*(6), e67374.

Gilmore, C., Attridge, N., & Inglis, M. (2011). Measuring the approximate number system. *The Quarterly Journal of Experimental Psychology*, *64*(11), 2099–2109.

Gilmore, C., Keeble, S., Richardson, S., & Cragg, L. (2015). The role of cognitive inhibition in different components of arithmetic. *ZDM*, *47*(5), 771–782.

Gilmore, C., McCarthy, S., & Spelke, E. (2010). Non-symbolic arithmetic abilities and mathematics achievement in the first year of formal schooling. *Cognition*, *115*(3), 394–406.

Graham, R. L. (1972). An efficient algorithm for determining the convex hull of a finite planar set. *Information Processing Letters*, *1*(4), 132–133.

Guillaume, M., Nys, J., Mussolin, C., & Content, A. (2013). Differences in the acuity of the Approximate Number System in adults: The effect of mathematical ability. *Acta Psychologica*, *144*(3), 506–512.

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "number sense": The approximate number system in 3-, 4-, 5-, and 6-year-olds and adults. *Developmental Psychology*, *44*(5), 1457–1465.

Halberda, J., Ly, R., Wilmer, J. B., Naiman, D. Q., & Germine, L. (2012). Number sense across the lifespan as revealed by a massive Internet-based sample. *Proceedings of the National Academy of Sciences*, *109*(28), 11116–11120.

Halberda, J., Mazzocco, M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–668.

Hanich, L. B., Jordan, N. C., Kaplan, D., & Dick, J. (2001). Performance across different areas of mathematical cognition in children with learning difficulties. *Journal of Educational Psychology*, *93*(3), 1–13.

Hasher, L., Zacks, R. T., & May, C. P. (1999). Inhibitory control, circadian arousal, and age. In A. Koriat & D. Gopher (Eds.), *Attention and performance XVII* (pp. 653–675). Cambridge, MA: MIT Press.

Hauser, M. D., Tsao, F., Garcia, P., & Spelke, E. S. (2003). Evolutionary foundations of number: Spontaneous representation of numerical magnitudes by cotton–top tamarins. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *270*(1523), 1441–1446.

Hellgren, K., Halberda, J., Forsman, L., Ådén, U., & Libertus, M. (2013). Compromised approximate number system acuity in extremely preterm school-aged children. *Developmental Medicine & Child Neurology*, *55*(12), 1109–1114.

Henik, A., Bibi, U., Yanai, M., & Tzelgov, J. (1997). The Stroop effect is largest during first trials. *Psychonomic Society*, *2*, 57.

Hiebert, J. (2013). *Conceptual and Procedural Knowledge: The Case of Mathematics*. Hillsdale, NJ: Lawrence Erlbaum.

Holloway, I. D., & Ansari, D. (2009). Mapping numerical magnitudes onto symbols: The numerical distance effect and individual differences in children's mathematics achievement. *Journal of Experimental Child Psychology*, *103*(1), 17–29.

Hurewitz, F., Gelman, R., & Schnitzer, B. (2006). Sometimes area counts more than number. *Proceedings of the National Academy of Sciences*, *103*(51), 19599–19604.

Hyde, D. C., Khanum, S., & Spelke, E. S. (2014). Brief non-symbolic, approximate number practice enhances subsequent exact symbolic arithmetic in children. *Cognition*, *131*(1), 92–107.

Inglis, M., Attridge, N., Batchelor, S., & Gilmore, C. (2011). Non-verbal number acuity correlates with symbolic mathematics achievement: But only in children. *Psychonomic Bulletin & Review*, *18*(6), 1222–1229.

Inglis, M., & Gilmore, C. (2013). Sampling from the mental number line: How are approximate number system representations formed? *Cognition*, *129*(1), 63–69.

Inglis, M., & Gilmore, C. (2014). Indexing the approximate number system. *Acta Psychologica*, *145*, 147–155.

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003a). Arithmetic fact mastery in young children: A longitudinal investigation. *Journal of Experimental Child Psychology*, *85*(2), 103–119.

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003b). A longitudinal study of mathematical competencies in children with specific mathematics

difficulties versus children with comorbid mathematics and reading difficulties. *Child Development*, *74*(3), 834–850.

Keller, L., & Libertus, M. (2015). Inhibitory control may not explain the link between approximation and math abilities in kindergarteners from middle class families. *Frontiers in Psychology*, *6*(685).

Kolkman, M. E., Kroesbergen, E. H., & Leseman, P. P. M. (2013). Early numerical development and the role of non-symbolic and symbolic skills. *Learning and Instruction*, *25*, 95–103.

Lee, K., Bull, R., & Ho, R. M. H. (2013). Developmental changes in executive functioning. *Child Development*, *84*(6), 1933–1953.

LeFevre, J. A., Fast, L., Skwarchuk, S. L., Smith Chant, B. L., Bisanz, J., Kamawar, D., & Penner Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development*, *81*(6), 1753–1767.

LeFevre, J. A., Smith Chant, B. L., Fast, L., Skwarchuk, S. L., Sargla, E., Arnup, J. S., ... Kamawar, D. (2006). What counts as knowing? The development of conceptual and procedural knowledge of counting from kindergarten through Grade 2. *Journal of Experimental Child Psychology*, *93*(4), 285–303.

Libertus, M. E., Feigenson, L., & Halberda, J. (2011). Preschool acuity of the approximate number system correlates with school math ability. *Developmental Science*, *14*(6), 1292–1300.

Libertus, M. E., Feigenson, L., & Halberda, J. (2013a). Is approximate number precision a stable predictor of math ability? *Learning and Individual Differences*, *25*, 126–133.

Libertus, M. E., Feigenson, L., & Halberda, J. (2013b). Numerical approximation abilities correlate with and predict informal but not formal mathematics abilities. *Journal of Experimental Child Psychology*, *116*(4), 829–838.

Libertus, M. E., Odic, D., & Halberda, J. (2012). Intuitive sense of number correlates with math scores on college-entrance examination. *Acta Psychologica*, *141*(3), 373–379.

Lindsay, D. S., & Jacoby, L. L. (1994). Stroop process dissociations: The relationship between facilitation and interference. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 219.

Lindskog, M., Winman, A., Juslin, P., & Poom, L. (2013). Measuring acuity

of the approximate number system reliably and validly: The evaluation of an adaptive test procedure. *Frontiers in Psychology*, *4*(510).

Logan, G. D., Zbrodoff, N. J., & Williamson, J. (1984). Strategies in the color-word Stroop task. *Bulletin of the Psychonomic Society*, *22*(2), 135–138.

Lonnemann, J., Linkersdörfer, J., Hasselhorn, M., & Lindberg, S. (2015). Symbolic and non-symbolic distance effects in children and their connection with arithmetic skills. *Journal of Neurolinguistics*, *24*(5), 583–591.

Lourenco, S. F., Bonny, J. W., Fernandez, E. P., & Rao, S. (2012). Nonsymbolic number and cumulative area representations contribute shared and unique variance to symbolic math competence. *Proceedings of the National Academy of Sciences*, *109*(46), 18737–18742.

Lyons, I. M., & Beilock, S. L. (2011). Numerical ordering ability mediates the relation between number-sense and arithmetic competence. *Cognition*, *121*(2), 256–261.

Lyons, I. M., Price, G. R., Vaessen, A., Blomert, L., & Ansari, D. (2014). Numerical predictors of arithmetic success in grades 1-6. *Developmental Science*, *17*(5), 714–726.

MacLeod, C. M. (1991). Half a century of research on the stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163–203.

Mazzocco, M., Feigenson, L., & Halberda, J. (2011a). Impaired acuity of the Approximate Number System underlies mathematical learning disability (dyscalculia). *Child Development*, *82*(4), 1224–1237.

Mazzocco, M., Feigenson, L., & Halberda, J. (2011b). Preschoolers' precision of the Approximate Number System predicts later school mathematics performance. *PLoS ONE*, *6*(9), e23749.

McLean, J. F., & Hitch, G. J. (1999). Working memory impairments in children with specific arithmetic learning difficulties. *Journal of Experimental Child Psychology*, *74*(3), 240–260.

Mejias, S., Grégoire, J., & Noël, M.-P. (2012). Numerical estimation in adults with and without developmental dyscalculia. *Learning and Individual Differences*, *22*(1), 164–170.

Mundy, E., & Gilmore, C. K. (2009). Children's mapping between symbolic and nonsymbolic representations of number. *Journal of Experimental Child Psychology*, *103*(4), 490–502.

Nigg, J. T. (2000). On inhibition/disinhibition in developmental psychopathology: Views from cognitive and personality psychology and a working inhibition taxonomy. *Psychological Bulletin*, *126*(2), 220–246.

Nosworthy, N., Bugden, S., Archibald, L., Evans, B., & Ansari, D. (2013). A two-minute paper-and-pencil test of symbolic and nonsymbolic numerical magnitude processing explains variability in primary school children's arithmetic competence. *PLoS ONE*, *8*(7), e67918.

Nys, J., & Content, A. (2012). Judgement of discrete and continuous quantity in adults: Number counts! *The Quarterly Journal of Experimental Psychology*, *65*(4), 675–690.

Odic, D., Hock, H., & Halberda, J. (2014). Hysteresis affects approximate number discrimination in young children. *Journal of Experimental Psychology: General*, *143*(1), 255–265.

Odic, D., Libertus, M. E., Feigenson, L., & Halberda, J. (2013). Developmental change in the acuity of approximate number and area representations. *Developmental Psychology*, *49*(6), 1103–1112.

Odic, D., Pietroski, P., Hunter, T., Lidz, J., & Halberda, J. (2013). Young children's understanding of "more" and discrimination of number and surface area. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(2), 451.

Park, J., & Brannon, E. M. (2013). Training the Approximate Number System improves math proficiency. *Psychological Science*, *24*(10), 2013–2019.

Parsons, S., & Bynner, J. (2005). *Does numeracy matter more?* London: National Research and Development Centre for Adult Literacy and Numeracy.

Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., ... Zorzi, M. (2010). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition*, *116*(1), 33–41.

Piazza, M., Pica, P., Izard, V., Spelke, E. S., & Dehaene, S. (2013). Education enhances the acuity of the nonverbal Approximate Number System. *Psychological Science*, *24*(6), 1037–1043.

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and approximate arithmetic in an Amazonian indigene group. *Science*, *306*(5695), 499–

503.

Price, G. R., Palmer, D., Battista, C., & Ansari, D. (2012). Nonsymbolic numerical magnitude comparison: Reliability and validity of different task variants and outcome measures, and their relationship to arithmetic achievement in adults. *Acta Psychologica*, *140*(1), 50–57.

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation? *Psychological Science*, *19*(6), 607–614.

Sasanguie, D., De Smedt, B., Defever, E., & Reynvoet, B. (2011). Association between basic numerical abilities and mathematics achievement. *British Journal of Developmental Psychology*, *30*(2), 344–357.

Sasanguie, D., Göbel, S. M., Moll, K., Smets, K., & Reynvoet, B. (2013). Approximate number sense, symbolic number processing, or number–space mappings: What underlies mathematics achievement? *Journal of Experimental Child Psychology*, *114*(3), 418–431.

Sasanguie, D., Van den Bussche, E., & Reynvoet, B. (2012). Predictors for mathematics achievement? Evidence from a longitudinal study. *Mind, Brain, and Education*, *6*(3), 119–128.

Shilling, V. M., Chetwynd, A., & Rabbitt, P. (2002). Individual inconsistency across measures of inhibition: An investigation of the construct validity of inhibition in older adults. *Neuropsychologia*, *40*(6), 605–619.

Smets, K., Gebuis, T., Defever, E., & Reynvoet, B. (2014). Concurrent validity of approximate number sense tasks in adults and children. *Acta Psychologica*, *150*, 120–128.

Smets, K., Sasanguie, D., Szűcs, D., & Reynvoet, B. (2015). The effect of different methods to construct non-symbolic stimuli in numerosity estimation and comparison. *Journal of Cognitive Psychology*, *27*(3), 310–325.

Soto-Calvo, E., Simmons, F. R., Willis, C., & Adams, A.-M. (2015). Identifying the cognitive predictors of early counting and calculation skills: Evidence from a longitudinal study. *Journal of Experimental Child Psychology*, *140*, 16–37.

Starr, A., Libertus, M. E., & Brannon, E. M. (2015). Number sense in infancy predicts mathematical abilities in childhood. *Proceedings of the National Academy of Sciences*, *110*(45), 18116–18120.

St Clair-Thompson, H. L., & Gathercole, S. E. (2001). Executive functioning as a predictor of children's mathematics ability: Inhibition, switching, and working memory. *Developmental Neuropsychology*, *19*(3), 273–293.

St Clair-Thompson, H. L., & Gathercole, S. E. (2006). Executive functions and achievements in school: Shifting, updating, inhibition, and working memory. *The Quarterly Journal of Experimental Psychology*, *59*(4), 745–759.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General*, *18*(6), 643–662.

Swanson, J., Deutsch, C., Cantwell, D., Posner, M., Kennedy, J. L., Barr, C. L., ... Spence, M. A. (2001). Genes and attention-deficit hyperactivity disorder. *Clinical Neuroscience Research*, *1*(3), 207–216.

Szűcs, D., Devine, A., Soltesz, F., Nobes, A., & Gabriel, F. (2013). Developmental dyscalculia is related to visuo-spatial memory and inhibition impairment. *Cortex*, *49*(10), 2674–2688.

Szűcs, D., Nobes, A., Devine, A., Gabriel, F. C., & Gebuis, T. (2015). Visual stimulus parameters seriously compromise the measurement of approximate number system acuity and comparative effects between adults and children. *Frontiers in Psychology*, *4*(444), 1–12.

Thorell, L. B., Lindqvist, S., Bergman Nutley, S., Bohlin, G., & Klingberg, T. (2009). Training and transfer effects of executive functions in preschool children. *Developmental Science*, *12*(1), 106–113.

Tronsky, L. N. (2005). Strategy use, the development of automaticity, and working memory involvement in complex multiplication. *Memory & Cognition*, *33*(5), 927–940.

Tzelgov, J., Henik, A., & Berger, J. (1992). Controlling Stroop effects by manipulating expectations for color words. *Memory & Cognition*, *20*(6), 727–735.

Vanbinst, K., Ghesquière, P., & De Smedt, B. (2012). Numerical magnitude representations and individual differences in children's arithmetic strategy use. *Mind, Brain, and Education*, *6*(3), 129–136.

Van Hoof, J., Janssen, R., Verschaffel, L., & Van Dooren, W. (2015). Inhibiting natural knowledge in fourth graders: Towards a comprehensive test instrument. *ZDM*, *47*(5), 849–857.

Wei, W., Lu, H., Zhao, H., Chen, C., Dong, Q., & Zhou, X. (2012). Gender

differences in children's arithmetic performance are accounted for by gender differences in language abilities. *Psychological Science*, *23*(3), 320–330.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1–B11.

Yeniad, N., Malda, M., Mesman, J., van IJzendoorn, M. H., & Pieper, S. (2013). Shifting ability predicts math and reading performance in children: A meta-analytical study. *Learning and Individual Differences*, *23*, 1–9.

Zhou, X., Wei, W., Zhang, Y., Cui, J., & Chen, C. (2015). Visual perception can account for the close relation between numerosity processing and computational fluency. *Frontiers in Psychology*, *6*(1364).