Dot comparison stimuli are not all alike:

The effect of different visual controls on ANS measurement

Sarah Clayton, Camilla Gilmore and Matthew Inglis

Mathematics Education Centre, Loughborough University, UK

Author Note

Sarah Clayton, s.clayton2@lboro.ac.uk; Camilla Gilmore, C.Gilmore@lboro.ac.uk; Matthew Inglis, M.J.Inglis@lboro.ac.uk; Mathematics Education Centre, Loughborough University, Loughborough, LE11 3TU, United Kingdom.

This research was supported by a Royal Society Dorothy Hodgkin Fellowship (C.G.), and a Royal Society Worshipful Company of Actuaries Research Fellowship (M.I.).

Correspondence concerning this paper should be addressed to Sarah Clayton, Mathematics Education Centre, Loughborough University, Loughborough, LE11 3TU, United Kingdom. Email: s.clayton2@lboro.ac.uk. Telephone: +44 (0)1509 22 8212.

The full data set and stimuli for this study are available to download at http://dx.doi.org/10.6084/m9.figshare.1546747.

Abstract

The most common method of indexing Approximate Number System (ANS) acuity is to use a nonsymbolic dot comparison task. Currently there is no standard protocol for creating the dot array stimuli and it is unclear whether tasks that control for different visual cues, such as cumulative surface area and convex hull size, measure the same cognitive constructs. Here we investigated how the accuracy and reliability of magnitude judgements is influenced by visual controls through a comparison of performance on dot comparison trials created with two standard methods: the Panamath program and Gebuis & Reynvoet's script. Fifty-one adult participants completed blocks of trials employing images constructed using the two protocols twice to obtain a measure of immediate test-retest reliability. We found no significant correlation between participants' accuracy scores on trials created with the two protocols, suggesting that tasks employing these protocols may measure different cognitive constructs. Additionally, there were significant differences in the test-retest reliabilities for trials created with each protocol. Finally, strong congruency effects for convex hull size were found for both sets of protocol trials, which provides some clarification for conflicting results in the literature.

Keywords: Dot comparison, Approximate Number System, Nonsymbolic magnitude comparison, Visual cues, Congruency effects, Numerical cognition

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

1. Introduction

The accuracy with which individuals can estimate numerical magnitude information has received increasing attention over the past decade. The Approximate Number System (ANS) is believed to be a cognitive system that underlies this sense of number (Dehaene, 1997). Many researchers have shown that individuals with superior performance on tasks designed to measure ANS acuity also demonstrate more advanced mathematical abilities (Halberda, Ly, Wilmer, Naiman, & Germine, 2012; Halberda, Mazzocco, & Feigenson, 2008; Libertus, Feigenson, & Halberda, 2011; Libertus, Odic, & Halberda, 2012; Mazzocco, Feigenson, & Halberda, 2011a; Piazza et al., 2010, Piazza, Pica, Izard, Spelke, & Dehaene, 2013; see Chen & Li, 2014 for a meta-analysis). This link has generated a great deal of interest from cognitive psychologists and education researchers, yet there are also many published studies which have failed to demonstrate this correlation (see De Smedt, Noël, Gilmore, & Ansari, 2013, for a review). The link between ANS tasks and mathematics and its implications have been widely debated, however previous research has given only limited attention to the development of the tasks assumed to provide a valid measure of ANS acuity. In this manuscript we explore how different protocols used by researchers to create dot comparison task stimuli may influence participant judgements and task reliabilities.

Several different tasks have been developed to measure ANS acuity, ranging from infant preferential looking change detection paradigms to more complex nonsymbolic arithmetic tasks (Barth, La Mont, Lipton, & Spelke, 2005; Xu & Spelke, 2000). A standard method of measuring ANS acuity in both children and adults is a dot comparison task. This task involves the brief presentation of two arrays of dots, usually on a computer, and requires the participant to select the more numerous array on multiple trials. The dot arrays are presented too briefly to count and so it is thought that individuals use ANS representations to determine which array contains the most dots (Dehaene, 1997). Performance can be measured by accuracy scores, reaction times or internal Weber fractions (referred to as '*w*' scores) which provide estimates of the acuity of approximate representations (Inglis & Gilmore, 2014; Price, Palmer, Battista, & Ansari, 2012). ANS acuities vary between individuals, and those with a more precise ANS are thought to generate numerosity representations closer to the actual numerosity and consequently perform better on nonsymbolic comparison tasks (Dehaene, 1997). Performance on ANS tasks is also dependent on the ratio between the two numerosities being compared; participants are more likely to make an error when comparing 19 vs. 20 dots than when comparing 10 vs 20 dots.

There is currently no universal procedure for dot comparison tasks and consequently different studies have used diverse methods of presentation. The dot array stimuli can either be presented simultaneously side-by-side (e.g. Gilmore et al., 2013), sequentially (e.g. Ansari, Lyons, van Eimeren, & Xu, 2007) or in an intermixed manner (e.g. Halberda et al., 2008). Dot comparison tasks can also vary by the number of trials used, the stimuli presentation times and the range of numerosities represented. Finally, and importantly for this study, there is no consensus on how the visual characteristics of the dot arrays should be controlled.

The stimuli in dot comparison tasks are produced with controls for visual properties which have the potential to bias responses to numerosity information. The visual characteristics of the dot arrays are therefore manipulated so that they are not consistently informative of number, i.e. the larger array is not always the more numerous. One common method of ensuring this is to control for the cumulative surface area of the dots. This is done by creating 50% of the task trials so that the numerically larger set is also larger in cumulative surface area, and 50% of the trials so that the numerically larger set is smaller (or sometimes equal) in cumculative surface area. This is the default setting on the Panamath software

(Halberda et al., 2008), and this type of control has been used in multiple studies of the ANS (e.g. Halberda et al., 2008; Halberda & Feigenson, 2008; Halberda et al., 2012; Hellgren, Halberda, Forsman, Ådén, & Libertus, 2013; Libertus et al., 2011; Libertus et al., 2012; Libertus, Feigenson, & Halberda, 2013a, 2013b; Mazzocco et al. 2011a; Mazzocco, Feigenson, & Halberda, 2011b; Odic, Libertus, Feigenson, & Halberda, 2013; Odic, Hock, & Halberda, 2014). In addition to the manipulation of cumulative surface area, Pica, Lemer, Izard and Dehaene (2004) also controlled for 'occupied area' often referred to as convex hull, or the total 'envelope area' taken up by the dot arrays. This method creates 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. Gebuis and Reynvoet (2011), developed this idea further to create a freely available online script whereby both cumulative surface area and convex hull are accounted for, but where trials could be 'partially congruent' in terms of either visual cue and numerosity. To elucidate, this method creates the following trials: 25% of stimuli pairs where the more numerous array also 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 and a smaller 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 but a larger convex hull. More recently researchers are starting to use this more comprehensive method of visual cue control (Defever, Reynvoet, & Gebuis, 2013; Gilmore et al., 2013; Inglis & Gilmore, 2013, 2014; Szűcs, Nobes, Devine, Gabriel, & Gebuis, 2013). Some researchers also report the average dot size and the density of the dots in the arrays, however these factors are highly correlated with cumulative surface area (Gebuis & Reynvoet, 2012), and so there is no substantial benefit to examining them as separate variables.

It is currently unknown whether the same skills underlie performance on all variants of dot comparison tasks. There have been some attempts to disentangle the cognitive demands and reliabilities of certain variations of the tasks (Price et al., 2012; Smets, Gebuis, Defever, & Reynvoet, 2014), however these have mainly focussed on the format of the tasks rather than the visual characteristics of the stimuli. Price and colleagues highlighted the lack of consistency in stimuli presentation between different tasks and report that the simultaneous presentation of the dot array stimuli produces the most robust effects. Sequential and intermixed presentation of stimuli may introduce extraneous cognitive demands such as increased working memory demands or the requirement to segregate visual information (Price et al., 2012). Inglis and Gilmore (2013) showed that differences in stimuli display times can influence performance on a dot comparison task. The longer an individual is given to process the information, the more precise the resultant ANS representation. This implies that different processes may be recruited to complete ANS tasks in which the stimuli are presented very briefly, to ANS tasks where the participant is able to view the stimuli until they respond. Consequently, it is difficult to meaningfully compare findings from studies that use different stimuli presentation times. Recently, Smets et al. (2014) have shown that different versions of nonsymbolic comparison tasks completed by the same participants provided significantly different ANS acuity estimates. They investigated concurrent performance on three versions of a dot comparison task: a standard dot comparison task, a same-different task in which participants were required to indicate whether two arrays represented the same or a different numerosity, as well as a change detection task in which participants were required to select which of two streams of changing arrays alternated between numerosities (e.g. 8-16-8 as opposed to 16-16-16). They found that the ANS acuity measures, both accuracy scores and *w* scores, obtained from each of the three tasks were not significantly correlated with each other.

Smets et al. (2014) conclude that there is a lack of validity across tasks that are all assumed to measure the ANS.

The finding that numerous different methodological factors can influence dot comparison performance is problematic for the development of research into the ANS. Many published studies have used diverse methodologies for nonsymbolic comparison tasks which render it difficult to build on previous findings. There have been some suggestions that *w* scores can be used to compare performance on dot comparison tasks that use different methodological formats (Piazza et al., 2013), but this is unlikely to be the case given Smets et al.'s (2014) results. Additionally, Odic et al. (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. Consequently, as also noted by Smets et al. (2014), conclusions gained from the comparison of *w* scores across different experiments may be flawed (e.g. Halberda & Feigenson, 2008; Piazza et al., 2010).

It is possible that variants of the task recruit the ANS to a different degree while also relying on varying levels of extraneous domain general processes. In particular it seems likely that inhibitory control is an important process involved in completing dot comparison tasks. There has been recent support for an inhibition account of performance whereby on certain incongruent trials of a dot comparison task, inhibition skills play a significant role (Fuhs & McNeil, 2013; Gilmore et al. 2013; Nys and Content 2012; Szűcs et al. 2013). Incongruent trials are those in which the more numerous array of the pair has smaller visual characteristics, such as cumulative surface area or convex hull. According to the inhibition account, participants may initially attempt to engage their ANS to judge which array is more numerous on all dot comparison trials, however, visual cues may interfere with this judgement on incongruent trials. For these trials, participants must inhibit the misleading visual cues of the stimuli in order to respond based on numerosity, making the task more difficult.

In direct support of the involvement of inhibition in dot comparison tasks, Cappelletti, Didino, Stoianov, and Zorzi (2014) have shown that older adults were particularly impaired on incongruent dot comparison trials, and that this impairment was correlated with poor inhibitory control as measured by a classic Stroop task (Stroop, 1935). Additionally, recent studies have found that the significant relationship between dot comparison performance and formal mathematics is moderated or even totally accounted for by inhibitory control skills (Fuhs & McNeil, 2013; Gilmore et al., 2013, Szűcs et al. 2013). Put together, it seems likely that some dot comparison task trials involve the recruitment of inhibitory controls skills, and it is possible that the way in which the visual characteristics of the stimuli are controlled may influence performance.

In line with the inhibition account, 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 et al., 2014; Gilmore et al., 2013; Hurewitz et al., 2006; Nys & Content, 2012; Szűcs et al., 2013). However, other studies have failed to find this effect (Odic et al., 2013, 2014), or find the opposite congruency effect (Gebuis & Van der Smagt, 2011). It is possible that this is partly due to the diverse methodologies for controlling visual cues employed in the tasks. For example, unlike the protocol designed by Gebuis and Reynvoet (2011), some methods do not allow researchers to systematically vary the convex hull of the dot arrays created. Notably, studies which have not found standard congruency effects did not manipulate convex hull size (Gebuis & Van der Smagt, 2011; Odic et al., 2013, 2014). It is therefore important to understand more about when congruency effects occur, and how both convex hull size and cumulative surface area influence task performance in one group of participants, across trials created with different visual cue controls.

This study aimed to investigate the reliability and concurrent validity of dot comparison tasks created using different stimuli protocols. We chose to examine the Panamath protocol, which has been widely used in ANS research and manipulates cumulative surface area, and the Gebuis and Reynvoet (2011) protocol which controls for both cumulative surface area and convex hull. We had 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 the convex hull size and the cumulative surface area of the dot stimuli influence accuracy? 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.

2. Method

2.1 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.

2.2 Task

Participants completed a nonsymbolic dot comparison task on a computer. 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.

There were 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 & Reynvoet (2011) protocol and block two consisted of 60 trials created with the Panamath protocol. 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 a block one first, and half completed block two first. Trials within the blocks were presented in a random order. Each trial began with a fixation point (600ms) followed by presentation of the two arrays (600ms) 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.

2.3 Stimuli

The Panamath protocol stimuli were downloaded from the Panamath website (http://www.panamath.org/9-12CollegeMaterials.zip; also 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[1] and numerosity (Figure 1). 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

[1] For the stimuli used in this study we found a high correlation between cumulative surface area and average dot size (*r* = .95) and density (*r* =.84), so for the remainder of the paper in our analyses we use only cumulative surface area.

cumulative surface area. 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) stimuli were generated using a freely available Matlab script provided online (http://titiagebuis.eu/Materials.html, Version May 20th 2011). This script controlled for cumulative surface area and convex hull, and generated four image types per trial (Figure 2). The first (fully congruent) included pairs of arrays where the more numerous array had a larger cumulative surface area and a larger convex hull. The second (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. Our intention 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. Specifically, the Gebuis and Reynvoet script contains a warning that "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 29-32 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. Consequently, in order to maximize comparability with existing literature, we used the Gebuis and Reynvoet (2011) protocol as close to its default setting as possible, ensuring that visual cues were controlled as intended. This involved choosing a slightly larger range of numerosities (22-36) within the typical range from the literature (Dietrich, Huber, & Nuerk, 2015). We chose to create 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 we chose not to alternate the side that each colour appeared as 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 1.

Table 1

Visual characteristics information for stimuli created with both 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 and convex hull.

Protocol	**Numerosity Range**	**Numerosity ratio range**	**Cumulative surface area ratio range**	**Convex hull ratio range**
Gebuis & Reynvoet (2011)	22-36	0.61- 1.64	0.10- 11.06	0.45- 2.35
Panamath (Libertus et al., 2012)	10-24	0.50- 2.00	0.34- 1.97	0.56- 1.60

3. Results

In the sections below we first present an analysis of the characteristics of the dot stimuli produced by each of the protocols. Then the relationship between performance on each of the protocol conditions and the test-retest reliability of the trials is explored using Pearson correlations. Finally, a by-items ANOVA was used to investigate the influence of convex hull and cumulative surface area congruency on participants' judgements. 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. This left 47 participants in the analysis.

3.1 Analysis of stimuli

We calculated, for each of the stimuli, the convex hull and cumulative surface area for the blue and yellow dot arrays. To calculate the convex hull for each array we used the Graham Scan algorithm (Graham, 1972). Cumulative surface area was calculated by summing the number of coloured pixels in the display. This allowed us to obtain measurements of each trial's visual characteristics using the same method for each protocol.

This analysis confirmed that the stimuli created using the Panamath protocol did not systematically control convex hull size, and therefore convex hull was predictive of numerosity on 37 of 60 trials. This is shown in Figure 3a 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 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 3b 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 evenly into each quadrant of the graphs shown in Figure 3c and Figure 3d for both protocols.

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 indicates that different processes may underlie performance on dot comparison tasks created with different visual controls.

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 Panamath protocol (60 in each block). To allow for comparability of reliabilities across blocks of trials created with these two different methods, we also calculated the test-retest reliability of a random subset of 60 Gebuis and Reynvoet protocol trials. We repeated this analysis 20 times, each with a different random subset of 60 trials to ensure the results were consistent. 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 higher than the Panamath test-retest reliability (*r* = .286).

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, we analysed congruency effects 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, 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 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. We explored this interaction 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 4) . 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 (*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 that cumulative surface area congruency.

For the Panamath trials, although we found convex hull congruency main effects that paralleled the Gebuis and Reynvoet trials, specifically higher performance on convex hull congruent in comparison to convex hull incongruent trials, we found a reverse effect for cumulative surface area congruency. Participants were more accurate on 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 5).

Table 2

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

Protocol	**Convex hull congruent**		**Convex hull incongruent**		**Cumulative surface area congruent**		**Cumulative surface area incongruent**	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Gebuis & Reynvoet	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

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 influence 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 was not significantly correlated. This 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), we 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 study when participants were re-tested with an average of 76.39 days between time one and time two, rather than immediately.

Our analysis of congruency effects replicates previous research (Barth et al., 2006; Gilmore et al., 2013; Hurewitz et al., 2006; Nys & Content, 2012; Szűcs et al., 2013) in demonstrating that performance on both Panamath and Gebuis and Reynvoet stimuli is influenced by the congruency status of dot comparison task trials, in particular the convex hull size. Moreover, our analysis demonstrates that accounting for the convex hull size as well as cumulative surface area is pivotal to understanding congruency effects. We have shown 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 smaller numerosity has a larger convex hull and larger cumulative surface area. This finding can provide 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 convex hull (e.g. Odic et al., 2013, 2014). The presence of convex hull congruency effects in trials created with both protocols provides support for the developing hypothesis that participants may recruit different mechanisms, such as inhibitory control skills, to complete incongruent trials (Fuhs & McNeil, 2013; Gilmore et al. 2013; Nys and Content 2012; Szűcs et al. 2013). Analogous results would not be found if we were to classify congruency 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 demonstrating 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 (Gebuis & Reynvoet, 2012). Given that there is much less range in the convex hull sizes of the Panamath stimuli, compared to the Gebuis and Reynvoet stimuli, our reverse congruency effect is in line with this finding. Our results therefore support Gebuis and Reynvoet's (2012) 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 present study adds 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.

Implications of our findings also apply to the controversial link between the ANS and mathematics achievement. 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.

To conclude, we have demonstrated that there is no correlation between adults' performance on dot comparison trials created by two protocols, which use different visual cue controls. Therefore divergent cognitive processes appear to underlie two nonsymbolic 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 cue 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 we have demonstrated that trials created with a widely used protocol have unacceptably low immediate test-retest reliability. In order to advance understanding of the ANS it may be necessary to focus future research efforts on continuing to develop nonsymbolic tasks which measure ANS acuity without the use of visual cues, such as auditory or tactile comparison tasks.

References

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,* 1845-1853.

Barth, H., La Mont, K., Lipton, J., Dehaene, S., Kanwisher, N., and Spelke, E. (2006). Nonsymbolic arithmetic in adults and young children. *Cognition, 98*, 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 of the United States of America*, *102*, 14116–14121.

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

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.

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*. Oxford: Oxford University Press.

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*, 48-55.

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).

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

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

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

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

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.

Graham, R. L. (1972). An efficient algorithm for determining the convex hull of a finite planar set. *Information Processing Letters, 26,*132-133

Halberda, J., & Feigenson, L. (2008). Developmental change in the acuity of the "Number Sense": The approximate number system in 3-, 4-, 5-, 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.

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.

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

Inglis, M., & Gilmore, C. (2013). Sampling from the mental number line: How are Approximate Number System representations formed? *Cognition, 129,* 63-69.

Inglis, M., & Gilmore, C. (2014). Indexing the Approximate Number System. *Acta Psychologica, 145*, 147-155.

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

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011a). Preschoolers' precision of the approximate number system predicts later school mathematics performance. *PLoS ONE*, *6*(9): e23749.

Mazzocco, M. M., Feigenson, L., & Halberda, J. (2011b). Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). *Child Development*, *82*(4), 1224-1237.

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 number discrimination in young children. *Journal of Experimental Psychology: General. 143*(1), *255-265.*

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

Pica, P., Lemer, C., Izard, V., & Dehaene, S. (2004). Exact and Approximate Arithmetic in an Amazonian Indigene Group. *Science, 306*(5695), 499–503.

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*, 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,* 1037-1043.

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.

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.

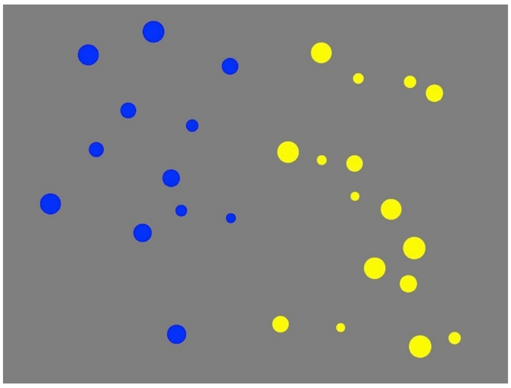
Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology: General, 18*(6), 643–662.

Szűcs, D., Nobes, A., Devine, A., Gabriel, F., & Gebuis, T. (2013). Visual stimulus parameters seriously compromise the measurement of Approximate Number System acuity and comparative effects between adults and children. *Frontiers in Psychology, 4*(444).

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

Figure 1. An example of a “correlated” (above) and “anti-correlated” (below) trial created with the Panamath protocol. Both images represent a 12 vs 16 dot trial.

1. Correlated trial



2. Anti-correlated trial

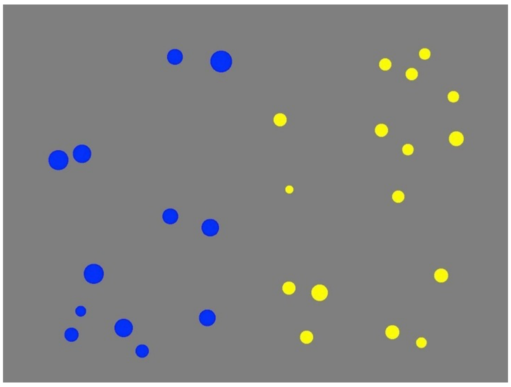
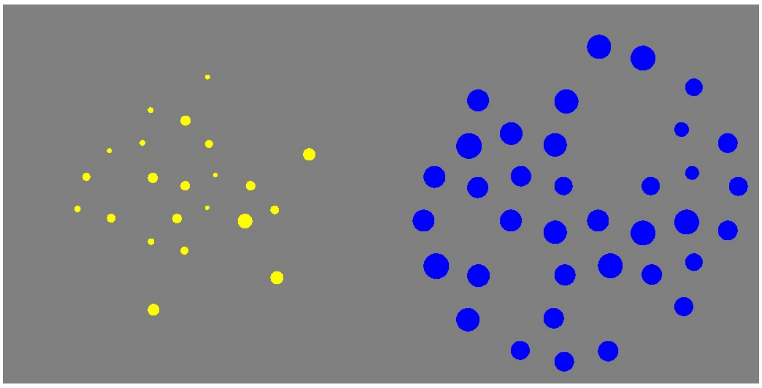
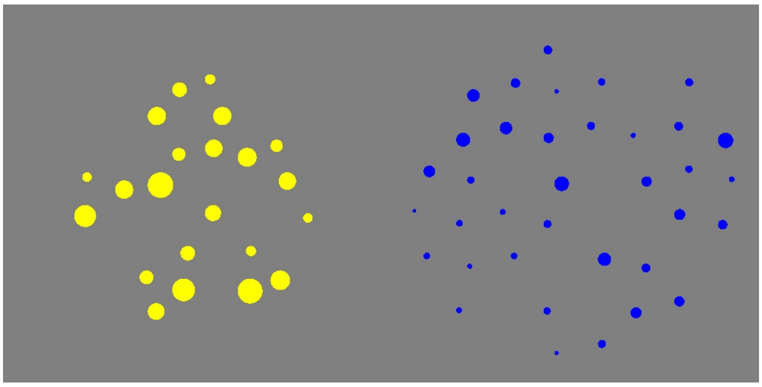


Figure 2. An example of the four image types created with the Gebuis and Reynvoet script. All images represent a 22 vs. 36 dot trial.

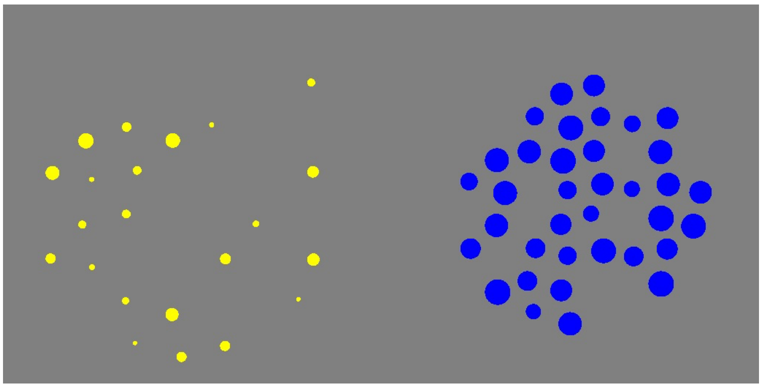
1. Fully congruent



2. Cumulative surface area incongruent, convex hull congruent



3. Cumulative surface area congruent, convex hull incongruent



4. Fully incongruent

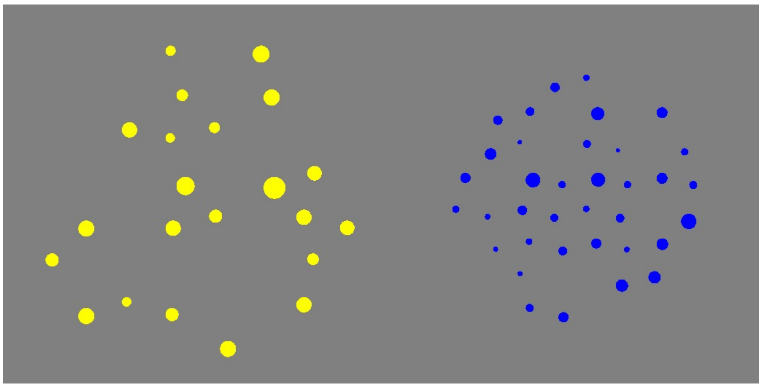


Figure 3. Dot comparison trials plotted in terms of the relationships between numerosity ratio and visual cue ratio for each protocol. (a) Numerosity ratio and convex hull ratio for Panamath trials, (b) numerosity ratio and convex hull ratio for Gebuis and Reynvoet trials, (c) numerosity ratio and cumulative 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.

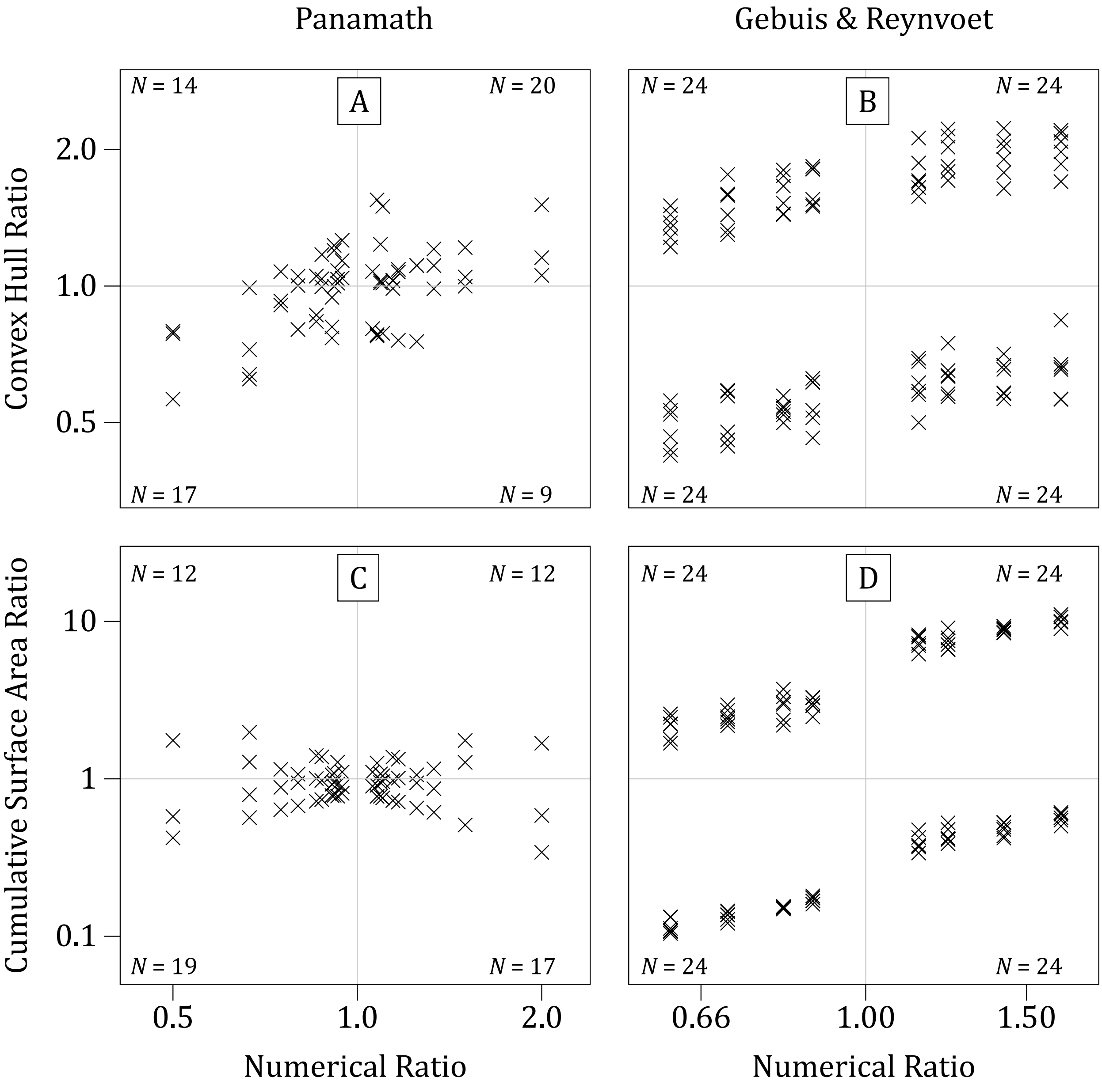


Figure 4. Interaction plot of mean accuracy scores on Gebuis and Reynvoet protocol trials calculated in terms of convex hull and cumulative surface area congruency.

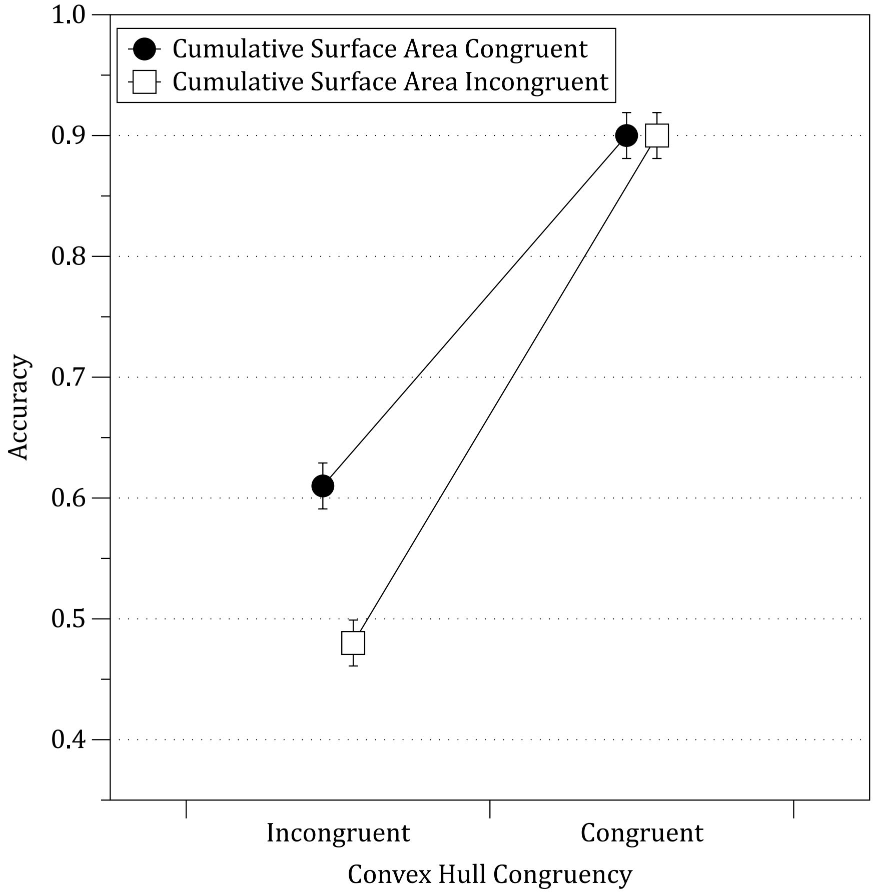
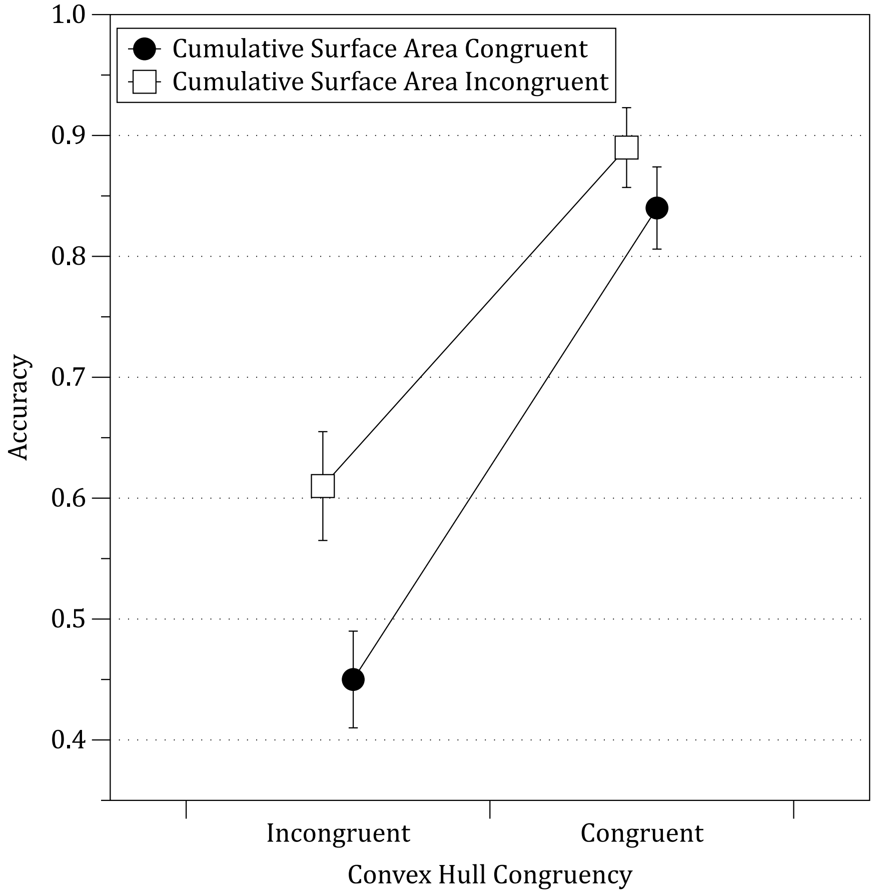


Figure 5. Interaction plot of mean accuracy scores on Panamath protocol trials calculated in terms of convex hull and cumulative surface area congruency.


● Cumulative Surface Area Congruent
□ Cumulative Surface Area Incongruent
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
Incongruent
Congruent
Convex Hull Congruency