Sussex Research Online

Do the colors of educational number-tools improve children's mathematics and numerosity?

Article (Accepted Version)

Rinaldi, Louisa J, Smees, Rebecca, Alvarez, James and Simner, Julia (2020) Do the colors of educational number-tools improve children's mathematics and numerosity? Child Development, 91 (4). e799-e813. ISSN 0009-3920

This version is available from Sussex Research Online: http://sro.sussex.ac.uk/id/eprint/84800/

This document is made available in accordance with publisher policies and may differ from the published version or from the version of record. If you wish to cite this item you are advised to consult the publisher's version. Please see the URL above for details on accessing the published version.

Copyright and reuse:
Sussex Research Online is a digital repository of the research output of the University.

Copyright and all moral rights to the version of the paper presented here belong to the individual author(s) and/or other copyright owners. To the extent reasonable and practicable, the material made available in SRO has been checked for eligibility before being made available.

Copies of full text items generally can be reproduced, displayed or performed and given to third parties in any format or medium for personal research or study, educational, or not-for-profit purposes without prior permission or charge, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

http://sro.sussex.ac.uk

Do the colors of educational number-tools improve children's mathematics and numerosity?

L.J. Rinaldi[1], R. Smees[1], J. Alvarez[1] & J. Simner[12]

[1]*School of Psychology, Pevensey Building, University of Sussex. BN1 9QJ. UK.*

[2]*Department of Psychology, University of Edinburgh, 7 George Square. EH8 9JZ. UK.*

Correspondence to: L.Rinaldi@sussex.ac.uk, School of Psychology, Pevensey Building, University of Sussex. BN1 9QJ

Acknowledgements

This research has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Grant Agreement no. [617678]. We are grateful to Duncan Carmichael, Monica Boutros and Olivia Casey-Haworth for their help in data collection and to John Maule for consulting on analysis. We declare no financial interest in conducting this research.

Abstract

This study examined how colored educational tools improve children's numerosity ('number sense') and/or mathematics. We tested children 6-10 years (n=3236) who had been exposed to colored numbers from the educational tools *Numicon* (Oxford University Press, 2018) or *Numberjacks* (Ellis, 2006), which map colors to magnitudes or Arabic numerals respectively. In a free-association task pairing numbers with colors, a subset of children spontaneously provided colors matching these schema. These children, who had internalized *Numicon* (colored magnitude), showed significantly better numerosity but not mathematics compared to peers. There was no similar benefit from internalizing *Numberjacks* (colored numerals). These data support a model in which colored number-tools provide benefits at different levels of numerical cognition, according to their different levels of cross-modal mappings.

Keywords: Numerosity, Mathematics, *Numicon, Numberjacks*, Color

Do colored number-tools improve children's mathematics and numerosity?

Early-years educators often use educational aids in mathematics, and these tools provide physical representations to make numbers more concrete (see Wing & Tacon, 2007). A large proportion of these tools pair numbers with colors, and these colored number-aids are aimed particularly at school children 4-11 years. For example, in one commonly used tool, *Numicon* (Oxford University Press, 2018), the numbers one to ten are physically represented as colored shapes with differing numbers of holes to represent magnitude (See Figure 1). The pairing of number with color and shape is assumed to promote mental imagery and this visualization of numbers is seen as key to the learning approach (see Wing & Tacon, 2007).

Tools such as *Numicon* are widely used in primary school education in the UK (Day & Lockwood, 2008; Devon Primary Math Team, 2006; Ewan & Mair, 2002; Wing & Tacon, 2007) as well as across Europe and in countries worldwide. The feature of interest in the current study is the color of these tools, since each number has an associated color which is consistent across all sets. Here, we investigated the degree to which the colors of tools such as *Numicon* may help children internalize numbers, and how this might impact on different numerical cognition skills. We look particularly at mathematics and numerosity ('number sense'; see below) and present a model predicting the efficacy of different colored number tools, in which colors bind to either magnitude or Arabic numerals and thereby influence different levels of numerical processing. We test our theory with data from over three thousand children who have been exposed to *Numicon* (encoding colored magnitude) or to a second tool which pairs colors to Arabic numerals (see below). We begin with a brief overview of the scientific literature on mathematical educational aids, and then introduce our model, hypotheses and study.

Figure 1. *Numicon* Shapes: A graphic representation of the *Numicon* shapes one to ten, which are individual 3-D plastic forms with the colors and configurations shown here.

Numicon represents just one example of the larger class of “math manipulatives”, an umbrella term for objects used in mathmatics to help make abstract numerical concepts more concrete (Clements & McMillen, 1996). These objects include not just *Numicon* shapes but also cubes, number rods, counting posters, and so on. Evidence suggests the use of these math manipulatives is a good pedagogical technique for teaching numeracy. For example, a meta-analysis by Carbonneau, Marley and Selig (2013) examined 55 studies comparing over 7000 students across the schooling years 6-18 years. Results showed that students using math manipulatives performed better than students using abstract symbols alone. Results were particularly striking in some areas over others, for example, with fractions showing a greater effect size than algebra or arithmatic. This suggests that math manipulatives can aid in different aspects of number cognition although the reasons for this are not entirely clear. Carbonneau et al. (2013) found the size of the effect was better when there was more emphasis on instruction given by educators, and was also influenced by age, with younger children showing moderate effects and older children showing smaller effects. However, this meta-analysis did not include *Numicon*, making it unclear how this particular math manipulative might fare.

Scientific validation for *Numicon* in particular (which we will use in our testing for the current study) has been attempted from a series of studies, many of which show numerical trends of improvements in mathematics for children using these tools, but often without statistical validation or control conditions (e.g., Tacon, Atkinson, & Wing, 2004; Ewan and Mair, 2002); Education Leeds, 2008; but see Nye, Buckley, & Bird, 2005). But the strongest case for support of *Numicon* was a randomized control trial by the *Education Development Trust* (Churches, 2016). This looked at two different math interventions including *Numicon*, within a sample of 875 low-performing students in School Years 1, 2 or 3 (between the ages 5-10 years). Approximately half the children were assigned to the *Numicon* group, and the rest were assigned to a control group where teachers continued teaching as they had before the study. Mathematical ability was measured before and after study using the *Progress in Math* tests which cover the UK mathematics curriculum (Clause-May, T., Vappula, H., & Ruddock, 2004). Churches (2016) found that *Numicon* was the only statistically successful intervention. In a replication one year later, controls and intervention children were assigned *within* each school to eliminate *a priori* differences across schools. Children in School Year 2 were randomly assigned to control or *Numicon* intervention groups and there was again a moderate but significant effect of improvement in the intervention group compared to controls.

A recent trend in the UK along with many other countries (e.g., USA) has been for more evidence-based policy (Gorard, See, & Siddiqui, 2017) and to the best of our knowledge, the study by Churches (2016) is the only randomized control trial on *Numicon*. The current study aims to contribute to the evidence-base on color-coded tools such as *Numicon* by taking a novel approach in investigating how and why *Numicon* might aid in number cognition, and placing these findings within a general theory accounting for the benefits of math manipulatives. We examined one specific aspect of *Numicon* in particular – its colors – to attempt to understand which features of *Numicon* might aid in which aspects of numerical processing and why. Tools such as *Numicon* have been colored deliberately on the assumption this plays a role in their educational benefits, so it is important to examine whether this choice has a meaningful effect. In our study, we took whole schools which have already been using *Numicon* and we looked across its pupils to find those who had internalized the *Numicon* colors. For this we ran a pre-test asking children to simply free-associate colors to numbers. We then measured how many times their color associations married with *Numicon* colors (e.g., the number five is red in *Numicon*; did they free-associate 5=red?). Comparing to chance levels, we took this as an index of whether *Numicon* colors had been mentally internalized by each individual child, and then used this metric to divide children into two groups: those who had internalized *Numicon* colors and those who had not. Finally, we took independent tests of numerical cognition across groups. If children had integrated the colors of *Numicon* into their mental number system, we asked whether, and in what areas, they might perform better in numerical cognition. Our theory predicts that improvements would be tied to the nature of the cross-modal coloring expressed by the manipulative itself (i.e., whether the tool associates colors with magnitudes or numerals, see below).

To understand our theory better, consider that we gave two types of numerical tests: a test of mathematics and a test of numerosity. Numerosity is our intuitive “number sense” which allows us to understand magnitudes without knowing the exact amount. This sense of numerosity relies on an *approximate number system* (ANS) which comprises a set of mental processes that approximately encode magnitudes (Dehaene, 2001). One common way of measuring numerosity is to ask participants to quickly discriminate between two arrays of dots, such as an array of black dots next to an array of white dots. Although the dots may be displayed too briefly to count, it is still possible to intuit whether the black or white dots were more numerous.

Adults are able to do this with great success (Barth, Kanwisher, & Spelke, 2003) and can discriminate dot arrays which differ by a factor of 1.15 or more (e.g., Lipton & Spelke, 2004). Even infants show evidence of an early ANS, and although their ANS is initially imprecise, it develops over time (Feigenson, Dehaene, & Spelke, 2004; Xu & Spelke, 2000). Importantly, the cross-modal nature of *Numicon* pairs color with – specifically -- magnitude (rather than numerals): its plastic shapes have pierced holes to represent magnitudes from one to ten and do not resemble numerals (see Figure 1). For this reason, we theorize that any advantage from the cross-modal influence of color would correspond to better performance in numerosity in particular.

We also included a comparison condition, which is a source that again matches numbers with colors, but this time pairs colors to Arabic numerals (rather than magnitudes per se). This baseline comes from a BBC television show (*Numberjacks*; Ellis, 2006) widely viewed by primary school children in the UK in which animated colored numerals 0 to 9 solve mathematical problems. The show was first released in the UK but has since been syndicated to countries worldwide, including the USA. This baseline allows us to test whether associating colors with numerals is beneficial in itself for processing magnitude, in which case a child who had internalized either *Numicon* or *Numberjacks* colors should show an advantage in numerosity. Alternatively, colored numerals may fail to benefit numerosity per se, because there is no cross-modal coding of color to magnitude itself.

Our model accounts for cross-modal advantages via the known benefits of "Dual-coding" (e.g., Clark & Paivio, 1991; Paivio, 1969). Here, colors would give number an enhanced level of encoding through a greater number of memory cues. These additional memory cues are assumed to strengthen representations and in term facilitate retrieval. In the case of *Numicon*, these additional cues are bound at the level of magnitude. In contrast, *Numberjacks* colors are bound at the level of Arabic numerals (but not magnitude directly), and this leads to our first prediction: Children who have internalized *Numicon* colors (dual-coding magnitude) should correspondingly have better performance in a test of numerosity, but these benefits would not be seen in children who have internalized *Numberjacks* (dual-coding numerals). Our second prediction is that children who have internalized *Numberjacks* colors might have better performance in our mathematics test, because this was designed around the UK curriculum and many of its questions are phrased using numerals.

Our third prediction comes from a consideration of how numerosity and mathematics interact. Importantly, although numerosity and mathematics ability are related (i.e., adults and children with high numerosity perform better in mathematics; Anobile, Stievano, & Burr, 2013; Chen & Li, 2014; Halberda Mozzocco & Feigenson, 2008), this relationship is assumed to be directional. Wong, Ho, and Tang (2016) have suggested a direction of causality in that a better ANS aids the ability to map numerosities to numerals and consequently improve math ability. This directional mapping from numerosity to mathematics suggests that children who have internalized *Numicon* colors (i.e., dual-coded magnitude) may perform better not only in numerosity but also in tests of mathematics. However, children who have internalized *Numberjacks* colors (i.e., dual-coded numerals) may not see similar benefits in numerosity. Our fourth prediction is that the known empirical relationship between numerosity and mathematics (improved numerosity correlates with improved mathematics; e.g., Anobile, Stievano, & Burr, 2013) is itself unrelated to color and will therefore operate irrespective of whether children have internalized colors from any device. Our model, and its four predictions are represented in in Figure 2 below.

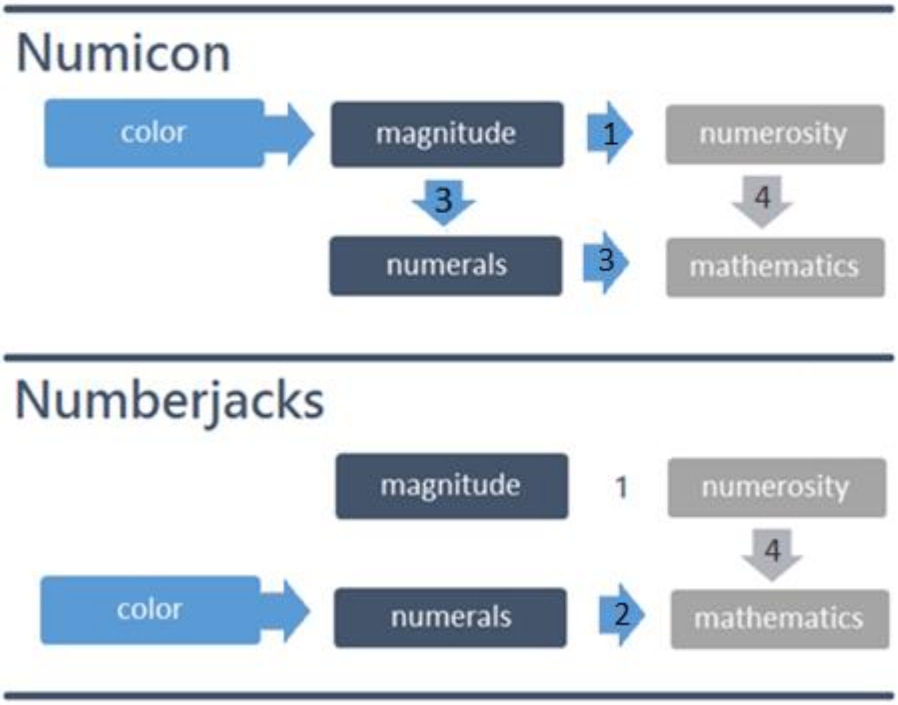


Figure 2. Modelling Math manipulatives: Left column shows the dual coding of color at different levels of representation from two types of math manipulatives (*Numicon* and *Numberjacks*). Middle column shows color is mapped directly to magnitude in *Numicon* but to numerals in *Numberjacks*. Final column shows the testing measures where our model predicts effects. Blue arrows show hypothesized dual-coding benefits from color, and grey arrows represent benefits unrelated to color. Our four hypotheses (see text) are mapped onto the model as numerals 1-4.

There have been very few studies of how numerosity (or indeed mathematics) might be improved by colored number tools like *Numicon* or *Numberjacks*. But in addition to the one study reviewed above (showing the efficacy of *Numicon* in mathematics), there is reason to think math manipulatives well might have an impact on numerosity. DeWind and Brannon (2012) showed that numerosity can indeed be improved with intervention: they trained 20 adults across six sessions on numerosity judgements with accuracy feedback and found that numerosity improved significantly. This suggests that the ANS is malleable so might be influenced by tools. Another line of evidence, this time relating to color in particular, comes from color-number associations in unusual populations. *Grapheme-color synesthesia* occurs in approximately 1.5% of adults (Simner et al., 2006; Carmichael, Down, Shillcock, Eagleman, & Simner, 2015) and children (Simner, Harrold, Creed, Monro, & Foulkes, 2009) and causes lifelong, automatic, quasi-idiosyncratic associations between colors and numerals (or between colors and letters/ words). There is a growing body of evidence that synesthetes perform better in certain cognitive domains (e.g., memory for words; see Meier & Rothen, 2013) and this has been linked by some to the same benefits of dual-coding we explore here (Radvansky, Gibson, & McNerney, 2011); hence synesthetes may have enhanced cognition because they dual-code graphemes with color information. We ask here, therefore, whether similar mechanisms of dual coding can also enhance cognition in non-synesthetes (and we return to this comparison with synesthesia in the *Discussion*).

Finally, regarding *Numberjacks,* there is some evidence that children do benefit from watching educational television: a longitudinal study by Wright et al. (2001) suggested preschool children had better receptive vocabulary, number skills and engaged in more reading if they had watched child-audience informative programs at ages 2-3 years. Furthermore, a meta-analysis of 24 studies in 15 countries suggested that children who watched the child-oriented show *Sesame Street* (Ganz Cooney & Morrisett, 1969-2018) performed better across basic literary, numeracy, science, health and safety, and pro-social reasoning (Mares & Pan, 2013). Together these results suggest that math manipulatives can influence learning, that benefits may come from colored numbers, that children can learn from television shows, and that both mathematics and numerosity show improvements from intervention. Finally, we point out that an alternative prediction is that colors might produce a *negative* effect on children's learning by increasing the cognitive load on mathematical thinking (e.g., McNeil, Uttal, Jarvin, & Sternberg, 2009). If colors become a distraction to learning they may inhibit numerical processing, or might directly inhibit processing certain types of math functions over others (e.g., inhibit arithmetic, where multiple colors could compete).

In summary, we present a study in which we elicited free-associations between colors and numbers from a group of over three thousand children. We used their responses to divide children into groups: those that had internalized *Numicon* colors versus those who had not; and those who had internalized *Numberjacks* colors versus those who had not. Finally, we tested whether children with internalized colors were better in tests of numerosity and mathematics. Our approach differs to previous studies in that we do not compare children according to whether or not they *use* tools such as *Numicon* (e.g. Churches, 2016), but we instead take a cohort who *all use these tools* and look instead at whether or not they have internalized the colors. Our key prediction is that those who had internalized *Numicon* (pairing color with magnitude) but not *Numberjacks* (pairing color with numerals) should correspondingly have better performance in a test of numerosity (i.e., magnitude judgements). A second prediction is that internalizing the colors of *Numberjacks* might be associated with increased mathematics performance by its direct color-coding of numerals. A third prediction is that internalizing the colors of *Numicon* may perhaps be associated with improved mathematics if any benefits from colored magnitude feed forwards into mathematics. A final prediction is that a relationship between numerosity and mathematics is also likely to exist independently of whether children have internalized colors.

Methods

Participants

In our Numerosity assessment we tested 3236 children aged 6-10 years (mean age =7.95; SD = 1.22). These were 1571 girls (mean age= 7.95, SD = 1.22) and 1665 boys (mean age= 7.95, SD = 1.22). Of these children, 92.5% were native English speakers. Children were recruited from 22 Infant and Primary schools across East and West Sussex in the south of England (n = 15 from East Sussex, n= 7 from West Sussex) and were in School Years 2-5 (for ages see Table

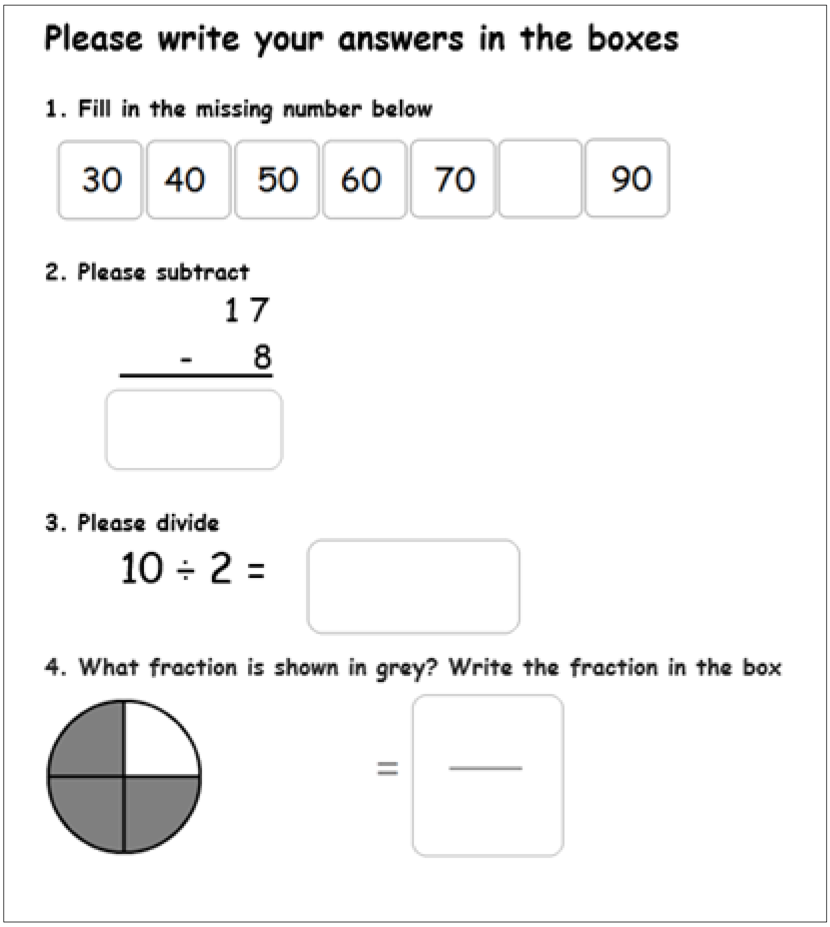
1). As an indicator of affluence/poverty (Taylor, 2018) the mean school-level free school meal (FSM) percentage was 13.44 %, where the national average from the same year is 14.5%, and our schools ranged in FSM status from 0.7% to 38.1%. In our Curriculum Math assessment, we tested a sub-group of these children, comprising n=2519 (mean age = 8.40; SD = 0.97; 1228 girls, mean age =8.39, SD =0.97; 1291 boys, mean age =8.40, SD = 0.96) who were children in School Years 3-5 only (see below for why Year 2 were tested for numerosity, but not curriculum math).

We also tested but excluded an additional 63 children. Of these, 20 were removed because they did not complete the tasks, and a further 33 experienced a technical error. Nine were flagged by teachers at our request as being newly arrived in the UK with particularly low levels of English, and one final child was out of year group (i.e., her chronological age did not match the rest of her class). Our study was approved by the local university ethics board and testing took place from October 2016 to the end of April 2017.

Materials and Procedure

Children were tested in class groups of approximately 30 within their classrooms, and they completed up to three tasks in the following order: a Curriculum Math test, a Numerosity test, and Colored numbers test. School Years 3-5 completed all tests, while Year 2 completed the latter two only because they had not yet covered enough of the math curriculum to be tested on mathematics (see below). Together our tests typically lasted for 5-10 minutes each but were interspersed with other activities (e.g., personality testing) to be reported elsewhere. These activities separated our tests by approximately 20 minutes.

Curriculum Math Test: we developed a short math test for children in School Years 3-5 based on the UK Primary school curriculum ("The national curriculum in England: Key stages 1 and 2 framework document," 2013). This test was presented on paper and there were 47 questions in total, one question for each of the 7-9 sub-sections of the math curriculum per years 1-6 (see example questions in Figure 3). For each child however, our test assessed knowledge of the curriculum for the child's current school year and two years prior. For example, Year 3 students start the test with questions from the Year 1 curriculum. (Since there is no set UK math curriculum prior to Year 1, students in Year 2 could not complete an equivalent test and were therefore excluded from mathematics testing). Children were given five minutes to complete as much as they could, as quickly and accurately as possible, and were not expected to go beyond their current year (e.g., Year 3 pupils start with Year 1 questions and, in general, are not expected to get to Year 4 questions). Children who got further than their current year were marked for all correct questions. Teachers and researchers gave no help to children, except with reading if necessary, and no feedback was given.



Please write your answers in the boxes

1. Fill in the missing number below

30	40	50	60	70		90

2. Please subtract

$$\begin{array}{r} 17 \\ -\ \ 8 \\ \hline \end{array}$$

3. Please divide

$10 \div 2 =$

4. What fraction is shown in grey? Write the fraction in the box

=

Figure 3. Mathematics Testing Materials. Example questions from our math test based on Year 2 (age 6-7) curriculum content.

Numerosity Task

The numerosity task (and the colored numbers task which follows) was presented on electronic tablets. Children were each given a touch screen Acer Aspire SW3-016 or Acer One 10 tablet, which ran on Intel® Atom TM x5-Z8300 Processors. These ran on Windows 10 and had 10.1" LED backlight touchscreens (1280 x 800 pixels).

Children were given the Panamath Numerosity Dot Task (Halberda et al., 2008) which required them to make a judgement based on dots on the screen. Children saw a cluster of white dots on the left side of the screen, and a cluster of black dots on the right side of the screen. Their task was to press one of two buttons (marked with a white or black sticker) to indicate whether there were more white dots or black dots. We used the default Panamath settings (Halberda et al., 2008) which generate an adjusted level of difficulty based on each child's age. The length of time for this task is adjustable and we set the task to run for two minutes. Children were told they would play a game in which they saw black and white dots on the screen. They were instructed to press a button to show whether there were more black or white dots. They were told they would not have time to count the dots so they should make their best guess as quickly as possible.

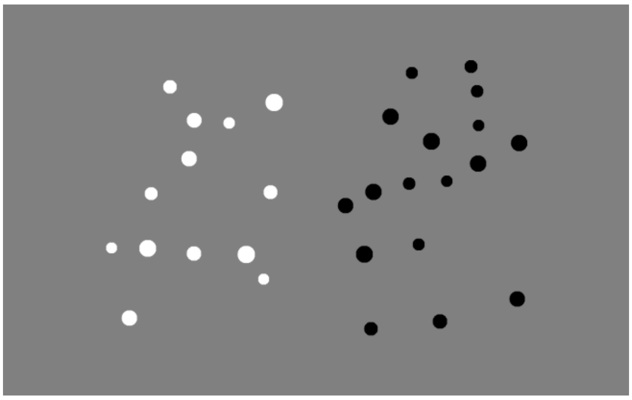


Figure 4. Numerosity Task. A screenshot of the numerosity dot task. Here the correct answer here is 'black' (i.e., screen shows 13 white dots and 17 black dots so black dots are more numerous).

Free Association: Number-Color Pairing Task.

We also tested children on a free association number-coloring task developed by our lab. In this test, children saw the numbers 0-9, individually in a random order, and were asked to think of the 'best' color for each number. Children chose their color using an on-screen color picker which appeared on the right-hand side of the screen and consisted of a vertical bar which could be dragged up and down to select hue. To the right of the hue bar was a 10x10 grid of color-swatches which allowed children to also select the exact shade. Within this 10x10 grid, luminance varied along one axis and saturation along the other, with axes randomly alternating from trial to trial. For example, if the child saw “5” and wanted to select a certain of red, (s)he would first drag the hue bar down to red, and would then inspect the 10x10 shade box to find the exact luminance and saturation of red required. Manipulating the color picker provided children with a choice from 25,600 discrete colors in RGB (red, green, blue) color space. Children were first trained to use the color-picker, which they managed without difficulty. Each trial began with a random initial setting. Graphemes appeared in lowercase black font, 2.5cm high, in a typeface suitable for children (Sassoon Infant®). Children inspected the number, which appeared on the left of the screen, and then made a color choice from the right-hand palette. Once the color choice was made, the program advanced to the next trial. (This task was also used to test an unrelated set of hypotheses to be reported elsewhere which required letters interspersed with the numbers, and presented three repeated blocks. For the present study, however, we report only the colors of numbers, taken from the first block only.)

Results

Participant exclusions

In our analyses, we examine two color schemas (*Numberjacks*, *Numicon*) with different exclusion criteria. For our *Numberjacks* analyses, there were no exclusions; i.e., all children were included since all were likely to have seen this extremely popular television show. For *Numicon*, we included only children whose teachers incorporated *Numicon* into their current teaching programs (since we could not otherwise guarantee children had seen these school-specific products). We therefore excluded 395 (non-*Numicon*) participants from our Curriculum Math analysis (this left n=2124 Years 3-5; mean age = 8.39; SD = 0.96) and the same 395 participants from our Numerosity analysis (3 of whom had already been excluded earlier for technical failures in Numerosity, see *Participants*; this left n=2844 Years 2-5; mean age = 7.89; SD = 1.23). (We point out that the same number of participants were excluded from both tests, despite more children taking Numerosity than Curriculum Math overall. This is simply because there were no exclusions among the extra (Year 2) children taking Numerosity). The year group and gender of participants within each analysis is shown in Table 1.

Table 1

Number of participants broken down by analysis, year group and gender

	Numberjacks		*Numicon*	
Year group	Female	Male	Female	Male
2	343	374	343	374
3	421	435	361	362
4	392	416	332	366
5	415	440	341	365
Total	1571	1665	1377	1467

Data preparation

In order to compare children's color choices with those from *Numicon* and *Numberjacks* we first coded children's color-choices into color categories (red, green, blue, yellow etc.). We next compared their chosen colors to those found within the comparison schemas of *Numicon* and *Numberjacks*. Details of this coding procedure are given below.

Color Categorization Coding.

For our color-categorizing, we took the RGB color space co-ordinates from each child's numbers 0-9 and transformed these into the 11 basic color categories of English (black, white, red, blue, yellow, green, orange, pink, purple, grey, brown; Berlin & Kay, 1991) using the following method. We based our categorizations on the XKCD color survey (Munroe, 2010) in which color co-ordinates within RGB color space were named with color-labels by 222,500 participants. We aimed to use XKCD color survey data to establish the boundaries in color space for each of the 11 basic color categories in English (e.g., what is the boundary of the color red? What is the boundary of the color blue? etc.). Once done, we would use these boundaries to classify children's color-coordinate into the 11 basic categories of English.

The participants of the XKCD color survey data (Munroe, 2010) named color co-ordinates using 949 color-terms (e.g., red, burgundy, pea green) so we first sorted these verbal color-labels into their 11 basic color categories. For this we were able to use definitions from the Oxford English Dictionary (OED) to classify 474 of these terms into either one color category (e.g., "navy" = blue), two categories (e.g., "violet" = blue + purple) or three categories (e.g., "violet pink" = blue + purple + pink). For the remaining 475 color-labels which had no clear OED definition, we recruited six researchers of color and sensory processing who were naive to the hypotheses of the experiment, to serve as coders. These coders were shown each color patch from Munroe (2010; which he had subsequently condensed by finding the central RGB of each repeated color-label using a stochastic hillclimbing algorithm). Coders were shown these patches on-screen alongside the names of 11 basic color categories. Coders were asked to simply select the best color category for each, and to select up to two categories where necessary. Coders agreed on all but 18 colors, and for these, both color categories were included (e.g., disagreement between blue and black resulted in both categories being accepted) which produced up to three color categories per item. This method provided us with boundaries in color space for each color category (red, green, blue etc.) which we could now apply to our children's RGB data. The outcome of our coding was that each child's color choice was now categorized within the 11 basic color terms of English.

Matching Colors to Schema.

Next we counted how many times each child had chosen a color for a number that matched either the *Numicon* or the *Numberjacks* schemas (see Table 2 below for the color-categories of these schema which were rated by two independent coders with 100% agreement). For example, if the child had chosen red for the number 5, this would count as a match for *Numicon* (whose 5 is red) but not for *Numberjacks* (whose 5 is blue). In cases where children's colors had been categorized as more than one color (e.g., a certain shade of turquoise was blue + green) a match was counted if either of the colors matched with the given schema. Each child received a single score for each schema (i.e., a *Numicon* score and a *Numberjacks* score) which was the total number of matches out of a maximum of nine for *Numicon* (1-9) and out of ten for *Numberjacks* (0-9).

We next established how many matches would constitute chance levels, using a Monte Carlo approach which simulated 10,000 children making free-associations between colors and numbers. Specifically, the simulation began with the 11 colors in English, which were *a priori* weighted to reflect how often they were chosen by children across our entire data set (e.g. blue was chosen more frequently than orange so was weighted accordingly). These weighted colors were then selected at random (with replacement) in ordered sets of nine (for *Numicon*) or sets of 10 (for *Numberjacks*). We repeated this 10,000 times and compared how each set matched to *Numicon* colors (or *Numberjacks* colors). In our *Numicon* simulation, for example, if the first color in the set matched the color of the *Numicon* shape for 1, this was a "match". If the second matched the color of the *Numicon* shape for 2, this is was another "match". This gave a match-score out of 9 for *Numicon* – and we did this repeatedly for 10,000 repetitions. This simply allowed us to establish the probability of matching to these schema by chance. Based on the conventional alpha of $p < .05$ we found the minimum number of matches to *Numicon* needed to exceed chance levels was five (and five matches was significant at $p = .011$). Five matches was also the appropriate statistical cut-off for *Numberjacks* (where five matches was significant again at $p = .011$). Given these analyses, we categorized children as using a *Numicon* or *Numberjacks* color-scheme if they had five or more matches (to *Numicon* or *Numberjacks* respectively), while children with four or fewer matches were considered to *not* be using these schema.

(A reviewer has asked us to also include an alternative approach where we identified internalizers against pure chance by running an equivalent Monte Carlo analysis but without weighting colors to reflect how often they were chosen by children; see Supplementary Information; SI. Either method identifies internalizers (at 5 matches as shown above, or 4 matches as shown in SI) and will produce exactly the same pattern of results in our subsequent analyses below. See SI Tables 3-6 for parallel analyses.)

Table 2

Color associations for numbers 0-9 in Numicon and in Numberjacks.

Number	*Numicon* color	*Numberjacks* color
0	n/a	green
1	orange	purple
2	blue	orange
3	yellow	pink
4	green	blue
5	red	blue
6	blue	yellow
7	pink	red

8	green	blue
9	purple	green

Within the total sample of children who had been exposed to *Numicon* (n=2844; Years 2-5), we found 26 children (0.9%) had internalized *Numicon*'s colors (i.e., they used *Numicon* as a coloring strategy more often than chance would predict). And within the total sample for *Numberjacks* (n=3236; i.e., all children; Years 2-5) we found 100 (3.1%) had internalized *Numberjacks* colors. Within School Years 3-5 only (i.e., the cohort for our Mathematics testing) these numbers were 1% (21 out of 2124) and 3.1% (78 out of 2524) respectively.

There was no overlap between *Numicon*-internalizers and *Numberjacks*-internalizers (as expected, since these use different color-schemes).

Modelling the influence of color schemas on Numerosity and Math ability.

Here we test the hypotheses that children who internalize *Numicon* colors may have better numerosity or curriculum math abilities than those who do not. Children's binary classification of using or not using the *Numicon* strategy to color numbers was entered into two hierarchical regression models to compare their performance first on the numerosity test and then on the curriculum math test. Our analyses will allow us to determine whether children who free associate the colors of *Numicon* are better in a test of numerosity and/or a test of mathematics (and we will then do similarly for *Numberjacks* colors).

Numicon in Numerosity. Our dependent measure, percent correct in numerosity, had a negatively skewed non-normal distribution. This was due in part to the nature of the score we used (percent correct, 50% being chance), and in part due to participants performing well on the task, so we took a bootstrapping approach in our regression model. (We did not use an alternative output from this test, a Weber fraction, because the Weber fraction cannot produce a score for children at or around chance level which is a valid score in our analysis.) Along with *Numicon* strategy (using or not using) we included chronological age as a predictor in step one because our data suggest that older children were significantly more likely to internalize *Numicon* colors than younger children (χ^2 (4) = 10.62, *p* = *.03*). Both predictors had a significant effect on numerosity: older children, and those who had internalized *Numicon* colors had better numerosity scores (see Table 3). The relationship between *Numicon* and numerosity equated to a gain of around 5% in numerosity scores for children internalizing Numicon colors. In order to further aid interpretation of this effect for *Numicon* colors, we investigated the Hedges g (quasi-equivalent to Cohen's d but for unequal groups). This effect size was small to moderate, Hedges *g* = 0.34. However, since this includes the influence of age, we re-examined Hedges g within a single age group (age 9, because this contained the largest set of Numicon internalizers) and produced a Hedges g = 0.37, suggesting the effect of Numicon is small-to-moderate.

Table 3

Numicon as a predictor of numerosity ability with 95% confidence intervals in brackets. Figures are based on 1000 bootstrap samples. Chronological was entered as years in decimals.

	b	SE B	*β*	*p*
Step 1				
Constant	72.40 (69.14 – 75.70)	1.65		.001
Age	1.66 (1.28 – 2.02)	0.19	.17	.001
Step 2				
Constant	72.46 (69.23 – 75.76)	1.65		.001
Age	1.65 (1.27 – 2.00)	0.19	.17	.001
Numicon Integration	3.10 (-0.33 – 5.87)	1.50	.03	.038

Note: R^2 = .03 for step 1; R^2 = .03 for step 2 R^2 change = .001

Numicon in Curriculum Math. Each correct answer on the math test was given a score of 1 and these were summed to generate an overall mark. These were converted to z-scores to allow us to compare scores across years given that children in different years saw different questions. We entered our z scores as the dependent measure in our regression model, along with age and *Numicon* strategy as predictors (with *Numicon* strategy as a dummy variable: ‘using strategy’ = 1 and ‘not using strategy’ = 0). Age was centered around the mean chronological age of the year-group, because each child received a test appropriate to his or her school year but could be older or younger *within the year*. Although age itself was a significant predictor of math ability (ß = 0.25, *p* <.001), the *Numicon* strategy (using or not using) was not (ß = .02, *p* =.465; see Table 4).

To explore this null result we performed a Bayes factor analysis to determine whether we have enough evidence to accept the null hypothesis (Dienes, 2014). Our Bayes assumes a half-normal distribution (Dienes, 2014) and we took our informative prior (i.e., a previous study against which to gauge our own findings) from a study showing improvement in math from a *Numicon* intervention (*Multi-sensory approach to the teaching and learning of mathematics. Pilot project 2005*, 2008). This chosen prior (unlike, say, Churches 2016) provided the statistical information necessary to calculate a Bayes Factor (i.e., a mean difference between groups that can be standardized, and the standard error of this mean). Bayes Factors lie on a continuum in which scores less than 0.33 constitute evidence for the null hypothesis, and scores above 3 indicate evidence for the experimental hypothesis (Dienes, 2014).Our moderate Bayes Factor (BF = 0.15) was indeed less than 0.33 allowing us to accept the null hypothesis (Dienes, 2014) with sufficient power to conclude there is no difference in math performance between children who had or had not internalized the *Numicon* color-system.

Table 4

Numicon as a predictor of mathematics ability with 95% confidence intervals in brackets. Chronological age is mean centered, within each school year.

	b	SE B	*β*	*p*
Step 1				
Constant	0.02	0.02		.451
Age	0.73	0.06	.25	<.001
Step 2				
Constant	0.02 (-0.02 – 0.06)	0.02		.412
Age	0.73 (0.61 – 0.85)	0.06	.25	<.001
Numicon Integration	-1.55(-0.57 – 0.26)	2.12	-.02	.465

Note: $R^2 = .06$ for step 1; $R^2 = .06$ for step 2

We end this section by pointing out that our pattern of results remains identical, even if we re-inserted all excluded children from our Numicon analyses. (These children had been excluded because they were not using *Numicon* in their current class, but were nonetheless likely to have been exposed to *Numicon* in younger years, given the schools we tested.) Figures 1-6 in our SI show histograms illustrating the number of matches to each schema for these participants, as well as our corresponding analyses re-inserting excluded children; our pattern of results remain the same (See SI- Tables 1 and 2).

Numberjacks in Numerosity. We turn now to the color-schema of *Numberjacks*. We performed the same regression analysis on our Numerosity data as above, but this time using the binary classification of whether children did, or did not color numbers according to *Numberjacks*. Our results show that age was again a significant predictor of numerosity performance, but *Numberjacks* was not (see Table 5). We again ran a Bayes Factor, here using an uninformed prior (in the absence of a suitable *Numberjacks* study for comparison) within Rouder and Morey's (2012) Bayes Factor calculator for regression models (found at http://pcl.missouri.edu/bayesfactor). Our JZS Bayes Factor was 0.05 which again is under .33 lending strong support for the null hypothesis (Lee & Wagenmakers, 2014).

Table 5

Numberjacks as a predictor of numerosity ability with 95% confidence intervals in brackets. Figures are based on 1000 bootstrap samples. Chronological was entered as years in decimals.

	b	SE B	*β*	*p*
Step 1				
Constant	72.99 (69.81 – 76.28)	1.61		.001
Age	1.57 (1.20 – 1.92)	0.18	.16	.001
Step 2				
Constant	72.98 (69.80 – 76.28)	1.61		.001
Age	1.57 (1.21 – 1.92)	0.18	.16	.001
Numberjacks Integration	0.12 (-2.50 – 2.34)	1.23	.002	.936

Note: $R^2 = .02$ for step 1; $R^2 = .02$ for step 2

Numberjacks in Curriculum Math. Finally, we repeated our analysis investigating whether the *Numberjacks* strategy (used or not used) in coloring numbers predicted math ability. Our results showed that age was again significant predictor for math but *Numberjacks* was not significant (see Table 6), and a moderate-to-strong Bayes factor of 0.09 confirmed our strong support for the null hypothesis.

Table 6

Numberjacks as a predictor of mathematics ability with 95% confidence intervals in brackets. Chronological age is mean centered, within each school year.

	b	SE B	*β*	*p*
Step 1				
Constant	-0.001	0.02		.968
Age	0.72	0.06	.24	<.001
Step 2				
Constant	0.001 (-0.04 - 0.04)	0.02		.952
Age	0.72 (0.61 - 0.83)	0.06	.24	<.001
Numberjacks Integration	-0.01 (-0.21 - 0.23)	0.11	.002	.905

Note: $R^2 = .06$ for step 1; $R^2 = .06$ for step 2

Relationships between Numerosity and Math ability.

It is important to note that numerosity skills usually correlate with math ability (Halberda et al., 2008), and for this reason we verified whether these also correlated within our own cohort. As expected, there was a significant relationship between numerosity scores and mathematics scores, such that children scoring highly on one measure were likely to score highly on the other ($r = .26$, $p < .001$). In other words, although internalizing *Numicon* colors predicted numerosity and did not predict math ability, there was nonetheless a significant relationship between numerosity and math (as we would expect, whether or not children had internalized colors).

Discussion

Our paper set out to investigate how colored numbering within the educational devices *Numicon* and *Numberjacks* might aid children's numerical cognition. We did this by identifying whether or not children had internalized the colors from each device, and the extent to which this aided them in a curriculum math test, and in a test of numerosity. To do this, we first asked children to free-associate colors to numbers and we inspected their responses to determine whether they had followed the schema of either *Numicon* or *Numberjacks* colors. We tested the math and numerosity skills of children who had internalized these colors, and compared them to controls who had not internalized either schema. Our model first predicted that *Numicon*, but not *Numberjacks*, would correspond to better performance on in numerosity (i.e., sense of magnitude) because *Numicon* maps colors directly to magnitudes, while *Numberjacks* maps only to numerals. Our data supported this prediction: children who had internalized the colors of *Numicon* performed significantly better in numerosity.

The second prediction from our model was that internalizing *Numberjacks* (colored numerals) might bring benefits in curriculum mathematics because many of our questions were numeral-based. This prediction was not supported. Relatedly we hypothesized that *Numicon* colors (which aid magnitude processing) might have a "knock on" effect in the same curriculum math test. Again this third prediction was not supported. We therefore conclude that within manipulatives such as *Numberjacks* which link color to numerals, the color itself does not aid in math ability. And that although manipulatives linking color to magnitude (*Numicon*) are associated with benefits in numerosity, the dual-coding of color to magnitude does not easily transfer to Arabic symbols – or may not help in mathematics even if it does so. Instead, our fourth hypothesis was supported: there was a significant correlation between numerosity and mathematics (see also Halberda et al., 2008) but this was not influenced by the colors of math manipulatives. We represent these findings in our updated model, shown in Figure 5. Although our model is phrased in terms of two particular math manipulatives (*Numicon* and *Numberjacks*) it makes generalizable predictions beyond these exemplars, and extends to any math manipulatives using color in its approach to teaching math.

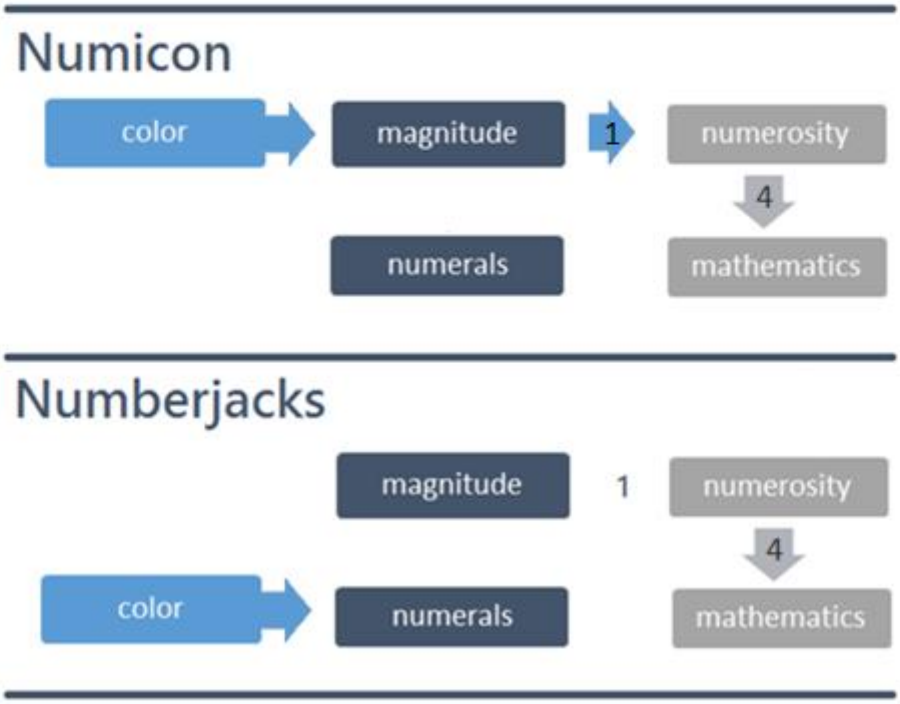


Figure 5. A Revised Model of Math manipulatives: Left column shows the dual coding of color at different levels of representation from two types of math manipulatives (*Numicon* and *Numberjacks*). Middle column shows color is mapped directly to magnitude in *Numicon* but to numerals in *Numberjacks*. Final column shows the testing measures where the model predicts effects. Blue arrows show hypothesized dual-coding benefits from color, and grey arrows represent benefits unrelated to color. Our data supported hypotheses 1 and 4.

Overall, our data suggests that an advantage in numerosity may come from Numicon’s colors. We have attributed this benefit in numerosity to *color* in particular because this was our independent variable (i.e., our groups were divided on whether they showed evidence of having internalized colors). Hence we have followed the standard empirical approach of attributing significance to the feature that was manipulated. However, it is logically possible of course (as in any study) that color may influence numerosity only indirectly, via some other correlating feature within Numicon (e.g., it could be that the shape of Numicon aids numerosity, and children who notice shape also happen to notice color). But there is no evidence in our study of any ‘middle-man’ influence, so we follow the assumptions of Occam’s razor in attributing advantages in numerosity to the internalization of Numicon’s colors, in particular.

So why might colored magnitude aid numerosity, but colored numerals *not* aid mathematics? And is this finding to be expected? A test-case for the impact of colors on numerical processing might be to re-examine whether benefits in math are seen in grapheme-color synesthetes. We saw earlier that synesthetes’ dual-coding of color with graphemes improves cognition (e.g., memory for words) but evidence within numerical cognition has been somewhat equivocal: Green and Goswami (2008; see Simner and Bain, 2018 for indepth analysis) found that three out of eight synesthetic children with colored numerals showed superiority in mathematics, with a group trend *p* =.09. But their recruitment methods could have encouraged high performing children irrespective of synesthesia (see Simner & Bain, 2018 for discussion). We are therefore in the process of administering a mathematics (and numerosity) test to approximately 50 synesthetic children whom we have recruited using random sampling methods. In summary, studies to date suggest synesthetic dual-coding of numerals may not bring unambiguous benefits in mathematics testing – which mirrors our findings here – and our future studies are exploring this further.

We should of course acknowledge the small number of children within our sample who will have had synesthesia – and even color vision deficiencies– although these will have made only a very small conitrbution to our study given our large sample. For example, given the low prevalence of grapheme-color synesthesia in the population (e.g., Simner et al., 2009) 99% of our sample will not have synesthesia, although some nonetheless demonstrated memory associations linking colors and numbers, as we have shown. Learned associations such as these can sometimes be difficult to distinguish from genuine synesthesia in a behavioral sense (e.g., Meier & Rothen, 2009) but they have different neurological correlates. Elias, Saucier, Hardie, and Sarty (2003) compared genuine number-color synesthesia against a case where colored numbers had been acquired from the environment (by a lifetime of cross-stitching, in which needles are colored and numbered). Although both cases performed similarly in behavioral testing, only synesthesia resulted in activation of the dorsal visual stream when manipulating numbers, suggesting synesthetes alone possess the quasi-perceptual phenomena that is unique to synesthesia. The case of acquired colors from cross-stitch needles is directly equivalent to our own cases here, suggesting that the children in our study who had internalized colors would likely be using similar, non-synesthetic neurological mechanisms.

It is important to clarify our claim that the colors from these math tools (*Numicon* and *Numberjacks*) were 'internalized' by some children. Our criterion was that children had to free-associate to the *Numicon* (or *Numberjacks*) colors more often than chance would predict. We assume that exceeding chance means that some psychological strategy was used, and this is the basis of our assumption that colors were 'internalized'. We point out that our 'internalizing' threshold of five or more matches to the nine colors of *Numicon* may seem small by intuition alone, but statistically-speaking this is highly improbable. And perhaps most importantly, our data show that 'internalizers' were indeed a meaningful cohort, because they were also the category who performed better in numerosity; i.e., this categorization had a detectable impact on scores. For this reason we are confident that our samples were meaningfully divided into children who have, or have not, internalized colors from being exposed to colored number-tools.

Although we found no effect of either *Numicon* or *Numberjacks* in our curriculum math test, it is important to point out this does not mean *Numicon* and *Numberjacks* do not improve math. Indeed prior studies testing *Numicon* would suggest otherwise (e.g., Churches, 2016). Here we can conclude only that math improvement in earlier studies is unlikely to stem from *Numicon*'s colors. It may therefore be the shape qualities of *Numicon* which improve math, or indeed some interaction between color and shape. And it is important to point out any limitations of our findings. We have assumed that *Numicon* colors improve sense of magnitude but the reverse might also be true: children with better numerosity ability may be better able to integrate colors into their magnitude schema. The nature of regression statistics do not allow us to infer the direction of causality, although findings elsewhere suggest that exposure to colored numbers in *Numicon* does causally induce changes in numerical cognition (Churches, 2016). We therefore tentatively assume in the absence of direct counter-evidence that the dual-coding of color aids in magnitude estimation rather than vice versa.

We point out that ours is the first study to examine the impact of *Numberjacks* on mathematical literacy and our results point to *Numberjacks* colors being influential at a surface level (e.g., children do internalize its colors) but not at a conceptual level (this did not lead to improvements in numerosity or math). One consideration, however, is that *Numicon* is actively taught at school, while *Numberjacks* is watched passively at home. Carbonneau et al. (2013) found that math manipulatives have an increased effect on children's learning when there is more emphasis on instruction given by the educator. We might therefore have found increased impact of colored numerals if these were used actively in the classroom, and we are now categorising schools according to their colored numeral displays (e.g., wall posters) in order to assess the impact this might have on math attainment.

We point out that only a small amount of variance was captured by our significant model (involving *Numicon* colors and numerosity) and this equated to a small-to-moderate Hedges *g* of 0.37. However, this effect size must be taken in context, and is almost certainly because we took an extremely indirect measurement of whether *Numicon* and *Numberjacks* had been internalized: we did not ask children for *Numicon*/ *Numberjacks* colors, and we did not mention *Numicon*/ *Numberjacks* to them in any way. We simply asked children to color numbers in any way they wished, but would likely have found a clearer influence of *Numicon* on numerosity if we had instructed children to recall *Numicon* colors directly. (We avoided this because we did not want children to think about math manipulatives in the context of our math/numerosity tests.) Nonetheless, even with our highly indirect measure, the interaction between *Numicon* and numerosity was significant and equated to a gain of around 5% in numerosity scores for children internalizing *Numicon* colors. Overall this suggests that internalising colored magnitudes might indeed aid in numerosity ability in a way that is important to acknowledge.

In conclusion, we found that some children internalize number-color associations from the educational tools *Numicon* or *Numberjacks*, which pair colors with magnitudes or Arabic numerals respectively. In the former case, we found a significant improvement in children's numerosity abilities, and conclude that *Numicon's* iconic representation of magnitude may help its colors become integrated into the ANS as a proxy for quantity. We found no benefits in mathematics testing, and no benefits in either numerosity or math for the colored numerals of *Numberjacks*. Together, our findings suggest that teaching magnitude-color patterns in education may be beneficial for the ANS in children's developing number cognition. Our findings would be of interest to a wide audience, including educationalists or researchers of developmental numerical cognition, or researchers of multisensory integration in learning, or indeed visual psychophysicists (we introduced a novel psychophysical metric for color categorisation). Finally, given that math manipulatives are common interventions for children with disabilities, our findings might also be relevant to clinical practitioners, and indeed to anyone interested in the benefits of internalizing environmental color. In summary, our results speak to the theoretical boundaries of multisensory learning, and to a fascinating interplay between numbers, colors and education.

References

Anobile, G., Stievano, P., & Burr, D. C. (2013). Visual sustained attention and numerosity sensitivity correlate with math achievement in children. *Journal of Experimental Child Psychology*, *116*, 380–391. https://doi.org/10.1016/j.jecp.2013.06.006

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults The construction of large number representations in adults. *Cognition*, *86*(3), 201–221. https://doi.org/10.1016/S0010-0277(02)00178-6

Berlin, B., & Kay, P. (1991). *Basic color terms: Their universality and evolution.* Univ of California Press.

Carbonneau, K. J., Marley, S. C., & Selig, J. P. (2013). A meta-analysis of the efficacy of teaching mathematics with concrete manipulatives. *Journal of Educational Psychology, 105*(2), 380–400. https://doi.org/10.1037/a0031084

Carmichael, D. A., Down, M. P., Shillcock, R. C., Eagleman, D. M., & Simner, J. (2015). Validating a standardised test battery for synesthesia: Does the Synesthesia Battery reliably detect synesthesia? *Consciousness and Cognition*, *33*. https://doi.org/10.1016/j.concog.2015.02.001

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. https://doi.org/10.1016/j.actpsy.2014.01.016

Churches, R. (2016). *Closing the gap: test and learn*. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/495580/closing_the_gap_test_and_learn_full_report.pdf

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review*, *3*(3), 149–210. https://doi.org/10.1007/BF01320076

Clause-May, T., Vappula, H., & Ruddock, G. (2004). *Progress in Mathematics 6*. GL Assessment.

Clements, D. H., & McMillen, S. (1996). Rethinking “Concrete” manipulatives. *Teaching Children Mathematics*, *2*(5), 270–279.

Day, J., & Lockwood, J. (2008). *Multisensory mathmatics for wave 3 intervention*.

Dehaene, S. (2001). Précis of “ The number sense .” *Mind and Language*, *16*(1), 16–36. https://doi.org/10.1111/j.1468-0017.1986.tb00086.x

Devon Primary Maths Team. (2006). *An image of number: The use of numicon in mainstream classrooms*. Exeter: Devon County Council.

DeWind, N. K., & Brannon, E. M. (2012). Malleability of the approximate number system: effects of feedback and training. *Frontiers in Human Neuroscience*, *6*, 68. https://doi.org/10.3389/fnhum.2012.00068

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*(July), 1–17. https://doi.org/10.3389/fpsyg.2014.00781

Elias, L. J., Saucier, D. M., Hardie, C., & Sarty, G. E. (2003). Dissociating semantic and perceptual components of synesthesia: behavioural and functional neuroanatomical investigations. *Brain Research. Cognitive Brain Research*, *16*(2), 232–237. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12668232

Ellis, C. (2006). BBC - CBeebies - Numberjacks.

Ewan, C., & Mair, C. (2002). Wiltshire Pilot Project - Numicon (March - July 2001). *Down Syndrome News and Update*, *2*(1), 12–13.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314. https://doi.org/10.1016/j.tics.2004.05.002

Ganz Cooney, J., & Morrisett, L. (n.d.). Sesame Street. Retrieved from http://www.sesamestreet.org/

Gorard, S., See, B. H., & Siddiqui, N. (2017). *The trials of evidence-based education: The promises, opportunities and problems of trials in education*. Taylor & Francis.

Green, J. A. K., & Goswami, U. (2008). Synesthesia and number cognition in children. *Cognition*, *106*(1), 463–473. https://doi.org/10.1016/j.cognition.2007.01.013

Halberda, J., Mazzocco, M. M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–668. https://doi.org/10.1038/nature07246

Kadosh, R. C., & Henik, A. (2006). Color congruity effect: Where do colors and numbers interact in synesthesia? *Cortex*, *42*(2), 259–263. https://doi.org/10.1016/S0010-9452(08)70351-4

Lee, M., & Wagenmakers, E. (2014). *Bayesian Cognitive Modeling: A practical Course.* Cambridge: Cambridge University Press.

Lipton, J. S., & Spelke, E. S. (2004). Discrimination of large and small numerosities by human infants. *Infancy*, *5*(3), 271–290. https://doi.org/10.1207/s15327078in0503_2

Mares, M.-L., & Pan, Z. (2013). Effects of Sesame Street: A meta-analysis of children's learning in 15 countries. *Journal of Applied Developmental Psychology*, *34*(3), 140–151. https://doi.org/10.1016/J.APPDEV.2013.01.001

McNeil, N. M., Uttal, D. H., Jarvin, L., & Sternberg, R. J. (2009). Should you show me the money? Concrete objects both hurt and help performance on mathematics problems. *Learning and Instruction*, *19*(2), 171–184. https://doi.org/10.1016/j.learninstruc.2008.03.005

Meier, B., & Rothen, N. (2009). Training grapheme-color associations produces a synesthetic Stroop effect, but not a conditioned synesthetic response. *Neuropsychologia*, *47*(4), 1208–1211. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/19350712

Meier, B., & Rothen, N. (2013). Synesthesia and Memory. In J. Simner & E. M. Hubbard

(Eds.), *The Oxford Handbook of Synesthesia* (pp. 692–707). Oxford: Oxford University Press.

Multi-sensory approach to the teaching and learning of mathematics. Pilot project 2005. (2008).

Munroe, R. (2010). Color Survey Results. Retrieved from https://blog.xkcd.com/2010/05/03/color-survey-results/

Numicon, Primary School Maths Resources. (2018). Retrieved from https://global.oup.com/education/content/primary/series/numicon/?region=uk

Nye, J., Buckley, S., & Bird, G. (2005). Evaluating the Numicon system as a tool for teaching number skills to children with Down syndrome. *Down Syndrome News and Update*, *5*(1), 2–13.

Paivio, A. (1969). Mental imagery in associative learning and memory. *Psychological Review*, *76*(3), 241–263. https://doi.org/10.1037/h0027272

Radvansky, G. A., Gibson, B. S., & McNerney, M. W. (2011). Synesthesia and memory: Color congruency, von Restorff, and false memory effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(1), 219–229. https://doi.org/10.1037/a0021329

Rouder, J. N., & Morey, R. D. (2012). Multivariate Behavioral Research Default Bayes Factors for Model Selection in Regression. *Multivariate Behavioral Research*, *476*(February 2013), 877–903. https://doi.org/10.1080/00273171.2012.734737

Simner, J., & Bain, A. E. (2018). Do children with grapheme-color synesthesia show cognitive benefits? *British Journal of Psychology*. https://doi.org/10.1111/bjop.12248

Simner, J., Harrold, J., Creed, H., Monro, L., & Foulkes, L. (2009). Early detection of markers for synesthesia in childhood populations. *Brain*, *132*(1), 57–64. https://doi.org/10.1093/brain/awn292

Simner, J., Mulvenna, C., Sagiv, N., Tsakanikos, E., Witherby, S. A., Fraser, C., … Ward, J. (2006). Synesthesia: The Prevalence of Atypical Cross-Modal Experiences. *Perception*, *35*(8), 1024–1033. https://doi.org/10.1068/p5469

Tacon, R., Atkinson, R., & Wing, T. (2004). *Learning about numbers with patterns. Using structured visual imagery (Numicon) to teach arithmetic*. London. Retrieved from https://www.numicon.co.nz/files/Learning_about_numbers_with_patterns.pdf

Taylor, C. (2018). The Reliability of Free School Meal Eligibility as a Measure of Socio-Economic Disadvantage: Evidence from the Millennium Cohort Study in Wales. *British Journal of Educational Studies*, *66*(1), 29–51. https://doi.org/10.1080/00071005.2017.1330464

The national curriculum in England: Key stages 1 and 2 framework document. (2013). Retrieved May 31, 2018, from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/425601/PRIMARY_national_curriculum.pdf

Wing, T., & Tacon, R. (2007). Teaching number skills and concepts with Numicon materials. *@BULLET Down Syndrome Research and Practice*, *12*(1). https://doi.org/10.3104/practice.2018

Wong, T. T.-Y., Ho, C. S.-H., & Tang, J. (2016). The relation between ANS and symbolic arithmetic skills: The mediating role of number-numerosity mappings. *Contemporary Educational Psychology*, *46*, 208–217. https://doi.org/10.1016/J.CEDPSYCH.2016.06.003

Wright, J. C., Huston, A. C., Murphy, K. C., Peters, M. S., Wright, J. C., Huston, A. C., … Kotler, J. (2001). The Relations of Early Television Viewing to School Readiness and Vocabulary of Children from Low-Income Families : The Early Window Project Marites Piñon , Ronda Scantlin and Jennifer Kotler Published by : Wiley on behalf of the

Society for Research in C. *Child Development*, *72*(5), 1347–1366.

Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), 1–11. https://doi.org/10.1016/S0010-0277(99)00066-9