A critical analysis of design, facts, bias and inference in the approximate number system training literature: a systematic review

Denes Szűcs, Timothy Myers

Correspondence: Denes Szucs; ds377@cam.ac.uk

Abstract: 133 words

A popular suggestion states that an evolutionarily grounded analogue magnitude representation, also called an approximate number system (ANS) or 'number sense' underlies human mathematical knowledge. During recent years many studies aimed to train the ANS with the intention of transferring improvements to symbolic arithmetic. Here we critically evaluate all published studies. We conclude that there is no conclusive evidence that specific ANS training improves symbolic arithmetic. We provide a citation analysis demonstrating that highly controversial results often get cited in support of specific claims without discussion of controversies. We suggest ways to run future training studies so that clear evidence can be collected and also suggest that data should be discussed in considering both supporting and contrary evidence and arguments.

Key words: number sense training; approximate number system; analogue magnitude system; numerical cognition; developmental dyscalculia; bias in research

Acknowledgments

The study was funded by the James S McDonnell foundation through the **James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition - Scholar Award** (DS; No 220020370).

Author contributions

DS designed the research, critically analysed the studies of interest and wrote most of the manuscript. TM carried out the literature search, the citation analysis, the description of the Number Race software and contributed to drafting.

1. Introduction

A popular suggestion is that an evolutionarily grounded analogue magnitude representation, also called an approximate number system (ANS) or 'number sense' underlies human mathematical knowledge (Dehaene, 1997). During recent years many studies aimed to train the ANS with the intention of transferring improvements to symbolic arithmetic. It is important to critically evaluate these studies because experience shows that interpretations are quickly taken up by researchers, practitioners and parents alike perhaps without much evaluation of how methods, results and study conclusions relate to each other, whereas usually the devil hides in the details. Unfortunately, many review papers tend to gloss over critical study details even though experimental design, analysis and/or inferential logic problems may inhibit clear conclusions or even disqualify results. Hence, in order to see clearly, here we critically review ANS training studies. We highlight both study-specific and general problems. We conclude that there is no conclusive evidence that specific ANS training improves symbolic arithmetic. We suggest ways to run future training studies so that clear evidence can be collected. We draw attention to the fact that highly controversial results often get cited in support of very specific claims in the literature without discussion of controversies. We suggest that this practice may facilitate the creation of a 'highly cited null field' which nevertheless gives an impression of positive results with regard to the ANS training literature. Below we first define important terms, then review studies one by one (because it is crucial to understand the details of individual studies so that they can be properly evaluated) and then draw some general conclusions. We especially point to the importance of bias-free discussion of results and placing them in the context of contrary as well as supportive literature.

1.1 What is number sense and the ANS?

A prerequisite of meaningful scientific debate is that we have a clear definition of what we wish to discuss. Literature regarding the ANS and number sense is often not up to this expectation as many researchers use this term in many different ways, and relevant definitions even seem to shift over time. Such confusions may result in some papers citing other papers as supporting evidence whereas they may have used completely different and non-compatible theoretical and/or operational definitions of number sense.

Here we assume that all the following terms mean the same: 'approximate number system', 'ANS', 'number sense', 'quantity representation', '(approximate) magnitude representation', '(approximate) analogue magnitude representation'. We take that all the above terms in the papers discussed below refer to the ANS in the sense defined by Dehaene (1997). This concept can be defined as an ancient, evolutionarily grounded pre-human sense of magnitude which represents numerosity (the number of items) in a modality-independent and approximate manner and it enables magnitude discriminations. Consequently, it is often claimed that this ANS is the intuitive pre-cursor of all human mathematics (Dehaene, 1997). It is to note that previously this concept was mostly called 'number sense', but more recently the tendency is to call it 'ANS'. It is also worth noting that the above ANS definition is very different from another popular, much broader, definition of 'number sense' which defines the term as a core set of early numerical abilities which are crucial to acquire for later numerical development to be successful (Jordan et al., 2006; Jordan et al., 2007; Jordan, Glutting, & Ramineni, 2010; Jordan et al., 2012; Hassinger-Das et al. 2014). This broader definition of 'number sense' includes both non-symbolic manipulation and symbolic counting and arithmetic principles. It assumes that number sense involves 1) magnitude comparison; 2) object and verbal counting; 3) number identification and 4) simple arithmetic. Here we only deal with the first definition of number sense or ANS. However, even a paper discussed here seems to blur the two definitions of number sense together (Sella et al. 2016).

Table 1 reviews the wide array of often *approximate* ANS definitions from the papers discussed here. Notably, several definitions provided do not necessitate an innate ANS and/or any special primitive representation of number. For example, the definition of Wilson et al. (2006b) could be satisfied by manipulating symbolic numerical quantities in visuo-spatial working memory by some spatial addition or subtraction algorithm. However, as far as we understand this would be an unintended extension of the definition of 'number sense' and ANS. Some other definitions are similarly imprecise (Wilson et al. 2009; Hyde et al. 2014), with probably DeWind and Brannon (2012) and Park and Brannon (2013) giving the most clear and specific definitions.

In the following, we will discuss each published study which can be thought of as aiming to train the ANS with the intention of demonstrating carry over (transfer) effects to other mathematical abilities beyond non-symbolic number comparison (see Appendix 1 for the method of identifying these studies). When we refer to tables and figures in the current paper we just give simple table and figure numbers. In contrast, we will use the '#' symbol when we refer to tables and figures in the actually discussed paper (e.g. **Fig. #7A** means Fig. 7A in the paper under discussion and not in this paper).

@ Table 1 about here

2. Training with the the Number Race software

Some studies used the so-called Number Race (NR) computer programme for training ANS (called 'number sense' or 'quantity representation' in these papers). For example, Wilson et al. (2009; abstract) states that 'The Number Race is an adaptive game designed to improve number sense.'. Wilson et al. (2006b) says that they define 'number sense' in a narrow way, as the term is usually used in the cognitive neuroscience literature (p2; bottom right; see **Table 1**). They justify the creation of NR by arguing that dyscalculia ('a disorder in mathematical abilities', 'due to specific impairment in brain function'; p2; top left) is a 'core deficit in number sense' (p3.) and argue that NR was designed with this 'core deficit in mind' (p4.). Here, they state that NR aims to provide 'intensive training on numerical comparison' and to emphasize the 'links between numbers and space' (p4.). However, while a focus on a supposedly 'core deficit' would assume fairly specific training, NR is a mixed bag of training interventions which may affect many other cognitive skills and representations besides the ANS.

NR instruction is built on three domains (Wilson et al. 2006a): First, it trains non-symbolic number comparison by prompting participants to choose between two groups of objects, one on the left and the other on the right. One of the two groups will have more objects than the other. For example, one group may have five objects while the other has three. There is also a timeline on the bottom of the screen with two characters, one for the player and the other representing the opponent. Whichever group the player chooses, the player's character will advance on the timeline the same number of spots as there were objects chosen and the opponent will automatically get the other group. So, if the player chooses the group with five objects his player will advance five spaces while the opponent would advance three. Since the first one to the finish line wins, it behoves the player to always try to choose the group with the most objects. The to-be-compared object arrays appear with varying levels of numerical distance between them, adapting to the comparison ability of the child. NR starts with easier number comparisons where there is large numerical distance between the to-be compared quantities and proceeds towards harder comparisons. The objects also appear in different sizes, either between or within groups. As will be shown below NR also aims to strengthen associations between spatial and numerical information. With regard to this, it is important to note that the ANS on its own is not supposed to include spatial elements, although this misconception is prevalent in the literature. In contrast, spatial-numerical associations seem culturally grounded (Dehaene, 1997), they appear gradually during development (e.g. White et al., 2012; Ebsersbach et al. 2008) and some researchers question whether they reflect properties of mature number representation at all, or they are rather related to working memory processes operating on representations (van Dijck & Fias, 2011).

A second domain that NR aims to train are links between various representations of number: non-symbolic representation, symbolic Arabic digits and aurally heard number words—primarily in that as the object arrays are shown, digits and aurally heard number words which correspond to the number of objects are also presented. This training domain goes well beyond the ANS: It constitutes both associative learning (linking representations) and training comparison operations with symbolic number representations. NR also presents the opportunity to practice a symbolic counting sequence. After the objects are transplanted from the top of the screen to the number line below, the narrator will name the spot which the player is at and then the avatar will jump a number of spaces to the new spot. While the spaces in between are not explicitly counted, the opportunity is there for the player to do so.

Third, NR also aims to increase the fluency of access to basic addition and subtraction facts. One way it does this is by stating the advancement of the player along the number line as an addition problem. For example, if the player is at spot 3 and chooses 5 objects, the programme will state, “Eight. Three plus five equals eight”. Sometimes the players will land on a trap. In this situation the programme will state the number of jumps back as a subtraction problem (e.g. “Oh no, you’ve landed on a trap. Eight minus two is six”.) Another way that arithmetic facts are reinforced is during the display of the two groups of objects. Occasionally the digits shown simultaneously with the groups of objects will be presented as an addition or subtraction problem. So, if the final number of objects is to be “four”, there might be six objects shown at first with the following symbolic expression: “6 – 2”. As the arithmetic expression is stated, two objects will simultaneously be separated from the group, visually showing four remaining. This trained domain may not have to do much with ANS at all as it is known that basic symbolic arithmetic operations are usually solved by memory retrieval processes (Ashcraft, 1982). In addition, the above training may also affect general visuo-spatial manipulations and visuo-spatial WM. While subtraction may rely more on quantity manipulation, NR of course cannot control whether these manipulations happen symbolically, by the use of a culture-specific mental number line, by relying on retrieved facts, or otherwise.

In light of the above it is very clear that NR affects much more than a putative core ANS system. This of course makes it hard to decide what exactly is being trained in studies using NR (as also acknowledged by Wilson et al. 2009; see later) which in turn makes the interpretation of results difficult. In fact, it is hard to see how NR is very different from some aspects of usual pre-school or school instruction which traditionally often uses concrete manipulatives (e.g. wooden blocks, counters, or Cuisenare rods; Boggan, Harper & Whitmire, 2010; Fuson & Briars, 1990; Hiebert, 1984; Marzola, 1987) to ground the concept of quantity. In addition, considering all the areas aimed to be covered by NR it seems that its range of trained activities is closer to the broad alternative definition of number sense used by Jordan et al. (2006; 2007; 2012) than to any focused definition of ANS. This ambiguity is also reflected in the fact that while the earliest NR studies exclusively discussed training as organized about a ‘core number sense’ (Wilson et al. 2006a,b; 2009), the latest NR training study (Sella et al. 2016) already seemed to define ‘number sense’ as trained by NR citing Jordan et al. (2012), whereas still citing some ANS studies as well. Hence, it seems that the implied definition of the core representations claimed to be trained by NR is *shifting*, converging on to Jordan et al.’s definition on number sense. Overall, while it is not clear what is being trained by NR we discuss NR intervention studies below as these are often cited in support of the role of ANS in mathematical development and symbolic math (see citation bias analysis in **Section 6**).

@ Table 2 about here

2.1 Wilson et al. (2006b)

Wilson et al. (2006b) used NR to train 9 seven to nine year-old children. There was no control group which disqualifies the study as a proper training study. The study also had very low power (**Table 2**) and one participant was even excluded from some analyses leaving only 8 participants in these. The training lasted for 5 weeks. Children trained 4 days a week for half an hour each day. Total training times ranged from 8 to 10 hours. Children were pre- and post-tested on an extended battery.

Comparing pre- and post-training test results showed that dot-enumeration (subitizing) performance became faster but its accuracy did not change (note that one participant was excluded from the dot enumeration analysis due to abnormal post-training data pattern). Non-symbolic number comparison accuracy and reaction time improved but the so-called ‘distance effect’ did not change. The distance effect (Moyer and Landauer, 1967) is considered one of the signatures of the ANS and it means that reaction times and error rates are larger in case of comparing closer as opposed to further away quantities. A change in ANS precision would imply a change in the distance effect. Such change was not detected by the study, so ANS precision did not change in response to training.

There was some improvement in subtraction performance. However, when the authors tried to qualify pairwise comparisons, none of the multiple testing uncorrected comparisons were close to the significance level (p=0.07 and p=0.08; p9). There was no improvement in symbolic number comparison and symbolic addition performance. It is interesting to note that the authors suggest that addition is a priori less related to quantity representation and manipulation than subtraction, so they state that they expected that addition performance will be unaffected by quantity training. However, many later ANS studies used non-symbolic addition for ANS training (Park and Brannon, 2013; 2014; Hyde et al. 2014) which suggests some confusion about this statement in the ANS literature.

The results from Wilson et al. (2006b) are inconclusive. First of all, outcome measures basically tested whether there was improvement on number skills *directly* trained by NR. Hence, it would not be surprising to see improvements. However, results were still inconsistent in that about half of trained domains did not show any improvement even after a 5-week intervention period, training 4 days a week. Of course, low power can be an explanation for this but at the same time low power can also be the reason that the study found relatively large overall improvement in post-training subtraction performance: It is well known that small, underpowered studies vastly exaggerate effect sizes because only occasionally atypically large deviations from a null effect are able to cross the statistical significance threshold when sample size is low (Button et al. 2013; see more on this later). Consequently, large effect sizes reported from underpowered studies generally cannot be trusted. Second, and most importantly, results are completely inconsistent with the follow up study's outcome discussed below (Wilson et al. 2009). This inconsistency can also be due to the low power of the 2006 study which increases the chances of false positive (random) statistically significant outcomes (Button et al. 2013). Third, the study was not up to even minimal standards of a training study since it did not have a control group. So, in principle the study should not be cited in support of any training claims as without a control group it is impossible to determine whether there were any NR-training specific effects. On the other hand, even if there were such effects, we could still not be able to determine what aspect of NR training exactly led to improvements due to the fuzzy nature of NR (see more below).

2.2 Wilson et al. (2009)

Wilson et al. (2009) divided 53 four to six year-old children into two groups. The paper does not say how many children were in each group but we may guess from the degrees of freedom (also note that t tests are communicated with 2 degrees of freedom, e.g. t[1,26]; which is incorrect). The 'math then reading' group (n = probably 27) was first trained with NR and afterwards with a reading training package, the other group vice versa ('reading then math' group; n = probably 26). There were pre-, mid and post-tests of number skills (time points T1-T3). The intervention happened during 14 weeks in 20-minute training sessions. NR training happened during 6 sessions, reading training happened during 4 sessions. This 6 vs. 4 session asymmetry was left unexplained. While the authors note in their Discussion that the mathematics and reading intervention time differed 'slightly' (p232; top left) the math intervention time was 150% expressed in the duration of reading intervention time which is more than a 'slight' difference. This discrepancy strongly biases the study for detecting a stronger effect of NR than reading training.

The statistical question was whether there would be a cross-over interaction of improved math performance between the 'math then reading' and 'reading then math' groups. To see this, separate pairwise comparisons for each group were also of interest between the T1-T2 and T2-T3 time points. These comparisons were multiple testing uncorrected t-tests.

While Wilson et al. (2006b) found that that NR improved non-symbolic comparison but not symbolic number comparison, Wilson et al. (2009) found just the opposite. This leads to questioning the results of both studies. In detail, Wilson et al. (2009) suggested that symbolic digit comparison improved specifically in response to NR training (although the T2-T3 contrast was n.s.; t(25)=1.69; p=0.1; which is equivalent to a small effect size: D=t/sqrt(26)=0.33; Fritz et al. 2012). In addition, NR also improved verbal symbolic numerical comparison. However, NR did not specifically affect the numerical distance effect, a marker of the ANS. Adding more negative findings, addition performance did not improve specifically in response to NR. Counting improved more in response to the control reading training than to NR. The authors note that this could be expected as counting is a more verbal operation and NR is not intended to train these. However, as noted above, NR may provide opportunities for counting as well (also see Räsänen et al. 2009; p.467; for a similar comment).

Strikingly, non-symbolic comparison performance also did not improve specifically in response to NR training and there was no change in the size of the numerical distance effect which is considered an important marker of the ANS. The lack of impact on non-symbolic comparison performance suggests that NR training does not affect at all the supposed core number sense skills it claims to improve—that is, its construct validity may be poor. Rather, thinking about the mixed nature of the NR package and the fact that in Wilson et al. (2009) NR improved symbolic number skills we may assume that it primarily trains general symbolic number comparison skills [Wilson et al. (2009) seems to have more credible findings as they had more power than Wilson et al. (2006b)].

The fact that symbolic number comparison performance improved but non-symbolic comparison performance has not is explained by assuming that NR improved 'number sense access' rather than 'number sense' per se. However, all we can observe is that accuracy and speed have improved on symbolic comparison tasks and the distance effect, an important marker of number sense has not changed in any tasks. So, actually nothing suggests that number sense played any role in improved performance. Rather, children just seemed to become faster and less error prone in working with Arabic digits and number words. A simple explanation would be that simply their symbol recognition and/or access to symbols per se has improved. There is no need to assume that number sense was involved in the observed findings. Rather, if number sense is important than we could also assume that its links with symbols were already strong enough before the intervention started and the intervention merely trained symbol access further. This is also likely because there was no change in the distance effect. For example, if the number/symbol links had become stronger due to training than we could have expected stronger activation of number sense, a consequence of which would most probably be a change in the distance effect. Such change was not observed. Hence, the authors' relatively ad hoc 'number sense access' improvement hypothesis is unnecessarily complicated and does not seem justified. Rather, it expresses interpretation bias for the number sense theory.

The authors provide a fairly unlikely explanation for some observed data, namely for the fact that NR specifically improved symbolic number comparison but it did not improve non-symbolic number comparison. They state that this would be so because their low socio-economic status participants' numerical problems were more related to access to number sense than to deficiency in number sense per se. That is, they *post-hoc assume* that their participants' number sense had no room for improvement while access to their number sense had room for improvement. In other words, the authors explain the null effect with regards to their most important training outcome by claiming that their participants had no need for this training but that the training was still efficient. This also implies that the authors take it for granted that NR has construct validity (trains what it claims to train) and rather conclude that their participants' otherwise unconfirmed internal properties explain their results. This inference interestingly combines tautology with a circular argument: If NR improves ANS we could assume that NR was successful. If NR does not improve ANS we assume that participants were in no need of improvement because otherwise ANS would have improved. That is, irrespective whether NR actually improved ANS or not, the authors always seem to be able to conclude that NR improves ANS by relying on some ad-hoc assumptions.

The authors also support their above argument by referring to their previous study (Wilson et al. 2006b) saying that it is unlikely that NR only improves 'access' to number sense (but not number sense per se) *'given previous results with the software in dyscalculic children… which showed improvement on non-symbolic as well as symbolic tasks'* (Wilson et al. 2009; p.231). However, in Wilson et al. (2006b) the improvement was detected in a subtraction task but in that study (as noted above) *symbolic* number comparison did *not* show any improvement in response to NR whereas *non-symbolic* comparison did. That is, the pattern of results regarding symbolic and non-symbolic number comparison was exactly the opposite between Wilson et al. (2006b) and Wilson et al. (2009) which obviously raises questions about the reliability of both findings. Simply put, first the authors first find 'A' but not 'B'. After this they find 'B' but not 'A'. Finally, they conclude that *both* findings 'A' and 'B' are valid. However, what they detected was a contradiction, or at least an inconsistency, rather than a confirmation of both 'A' and 'B' (because they *failed* to replicate their findings). The paper's argument is misleading because there is no mention of this crucial inconsistency. In addition, the power of Wilson et al. (2009) was much higher than that of Wilson et al. (2006b). Because small studies are likely to produce false positive findings with large effect sizes (Schmidt 1992) it is likely that the findings of Wilson et al. (2009) are to be trusted more than the findings of the earlier study. This would mean that the authors' explanation can be discarded to start with.

Finally, as discussed above, NR aggregates non-symbolic and symbolic numerical training, fact retrieval training, developing spatial-numerical associations, and even offers counting opportunities (Räsänen et al. 2009). Hence, it would be unjustified to state that any improvements would be related to training number sense (ANS) or 'number sense access'. The authors are conscious of this as they say: 'The present work… suffers from the difficulty of pinpointing precisely which instructional feature is responsible for the effect found.' (Wilson et al. 2009; p232; top right). Then they also note that they think that the 'improvement is in number sense access rather than in number sense per se' and just above they conclude that NR 'can be used for targeted instruction of number sense' (p232; bottom right). The authors finally conclude that 'although a targeted cognitive intervention such as our software is not intended to replace large-scale curricular interventions, it carries several benefits' (p233). This is an unjustified statement after not being able to show improvement—neither on the supposedly most important 'core component' of the tasks nor in addition, as well as delivering inconsistent results with their own previous study. Moreover, counting, an important school instruction outcome improved more in response to the control intervention with 50% less training time than NR training.

So, while the statement above is technically correct (NR '*can* be used for instruction') it would also imply that this instruction would be *specific* and *successful*. However, the study provides evidence of neither of these claims/implications: 1) Non-symbolic and symbolic comparison results are inconsistent across Wilson et al. (2006b) and Wilson et al. (2009). 2) There has been no change in the distance effect, an important marker of the ANS. 3) The studies cannot determine what exactly was trained because of the fuzzy nature of NR.

2.3 Räsänen et al. (2009)

Räsänen et al. (2009) chose a more optimal design than the above studies. They had two training groups each consisting of 15 mathematically underachieving 6.5 year-old children and an unseen control group of 29 children. One training group used NR, the other used another game aiming to improve mathematical skills, called GraphoGame-Math (GG) (Mönkkönen et al. in preparation). The two games have different approaches to improving early math skills. NR was built with the ANS in mind and therefore it starts with emphasizing approximate comparisons of relatively distant numerosities (note, however, that NR also verbally reads numbers while showing objects and visual digits from the early stages of the game). GG aims to start in the opposite way, emphasizing small sets of similar numerosities and to build more on verbal mediation. The main question was which game would improve math outcomes more. Children trained for 10-15 minutes per session, once a day for 3 weeks.

GG training lowered children's reaction time in symbolic number comparison in a larger extent than the improvement measured in the control group. NR training also seemed to have similar impact but none of the pairwise comparisons were significant ($p=0.069$ and $p=0.061$). None of the games resulted in any improvement in any other areas of number skills tested (verbal counting, object counting, 3-minute paper and pencil addition and subtraction test). These results are in conflict with Wilson et al. (2006b; 2009).

First, it is worthwhile to mention that even the unseen control group showed steadily decreasing reaction times from the pre- through the post-test and a delayed post-test. Hedges' G (hereafter 'D') can be computed from data in Table #3 for within-group differences (see **Appendix 1**). The improvement of the unseen control group from pre to post-test was $D=0.13$ and from pre to delayed post-test was $D=0.26$ even if this group did not have any intervention. These data have general significance and strongly suggest that noticeable improvement can happen with the passing of time in young children even when they do not have any targeted instruction.

Second, we can also compute within-group pre to post and pre to delayed post-test improvement for GG and NR. These values are $D=0.64$ (pre to post) and $D=0.88$ (pre to delayed post) for GG and $D=0.31$ and $D=0.45$ for NR. The picture is similar when effect sizes are computed comparatively between intervention and control groups (see Table #4). At both post-test and at delayed post-test GG achieved about 40% larger effect size compared to NR (0.52/0.36 and

0.53/0.38). Hence, GG achieved much larger speed improvement in terms of standardized effect sizes than NR (also, and as noted above, NR time contrasts were n.s.). This strongly suggests that GG was superior to NR when it comes to training number comparison reaction time.

2.4 Obersteiner et al. (2013)

Obersteiner et al. (2013) had a similar goal to Räsänen et al. (2009) and aimed to compare the impact of ANS-based approximate training with more exact number training. They developed two versions of NR. The approximate training focused on approximate number comparison, estimation and calculation. Time pressure in the game aimed to make sure that participants rely on approximate strategies. In the exact game version participants had to match numbers exactly. This was achieved by presenting alternatives differing only by 1 unit.

147 children were divided into 4 training groups. One group received approximate training (n=35); one exact training (n=39), one mixed training alternating session by session (n=39). A control group used a language training software (n=34). Each child in each group took part in 10 training sessions for a duration of 30 minutes over a period of 4 weeks.

Reaction time served as the outcome measure. Approximate training improved speed in non-symbolic and symbolic magnitude comparison and in approximate calculation. Exact training improved speed on a canonical subitizing task where dots were arranged in patterns. Subitizing with random dot patterns was not improved by any of the trainings. Both the exact and approximate training seemed equally effective in improving math test outcomes, both overperforming the combined training. However, effect sizes seemed very small in relation to standard errors (see Table #4). Indeed, relevant post-hoc comparisons contrasting the control group with the approximate and exact training groups were not significant (p=0.059; p=0.057). The authors opined that the observed improvements were small (p.133), there was only about a 2 score range of math outcome scores across the 4 groups (see Table #4) whereas the full range of scores was 0-45. Overall, the authors concluded that both exact and approximate tasks only improved performance on tasks which included exact and approximate components and both generated small gains on the mathematical test.

The original paper only communicated post-test scores from ANCOVAs where pre-test scores were taken as covariates. Therefore, we reanalysed post-test minus pre-test score differences by means of computing bias corrected and accelerated 95% bootstrap confidence intervals (100,000 permutations) for score changes from pre to post-test (We are grateful for the generosity of Andreas Obersteiner for providing us the data.). Consistent with the original report we found that 95% confidence intervals strongly overlapped for the approximate (2.91 – 6.14) and exact (1.77 – 5.74) groups and the mixed (1.13 – 3.72) group also had overlapping intervals with the above two groups. The no training group (-0.26 – 3.44) was also close to not having zero value in the confidence interval and showed definite overlaps with all training groups (this is consistent with the fact that the original paper only found marginally significant differences between the training groups and the control group). Hence, in line with the authors we conclude that there were no clear differential effects of the interventions and any training effects were small. In addition, similarly to the data of Räsänen et al. (2009) there were indications that even the unseen control group may have improved somewhat which again directs attention to the fact that children's performance may improve even if they are not trained.

2.5 Sella et al. (2016)

Sella et al. (2016) tested 5-year-old children and in principle assigned 23 children to NR training and 22 children to a control training. However, less (sometimes many less) children's data were analysed, so power varied greatly (**Table 2**). Children had at most two 20 minute-long activity sessions per week for 10 weeks (on average 16.9 NR sessions vs. 16 control sessions). The main flaw of the study is that the control training was not a properly designed training activity but an unstructured drawing program where 'kids [were] presented with a blank canvas and a variety of drawing tools to help them be creative' (www.tuxapaint.org; quote from the website; retrieved 27 June 2016). So, first, the control activity did not provide the same level of intellectual enhancement: As the authors themselves note 'NR was a more meaningful activity' than the control activity (p.27).

Second, the control activity was not a well-matched control training if mathematical improvement was of any interest because no impact of it can be expected on mathematics a priori (see more below).

The consequence of the above design problems is that there was a huge imbalance in mathematical instruction received by the training and the control groups. The authors note that the 'regular scholastic program' received by both groups 'entailed numerical activities for half an hour once a week' (p.23). That is 30 mins per week for 10 weeks, 10×30=300 minutes of regular mathematics instruction for both groups. On top of this the training group received about 16.9 sessions × 20 minutes = extra 338 minutes of extracurricular targeted mathematics instruction through NR while the control group was taking part in unstructured drawing. So, the training group received more than twice as much mathematics instruction than the control group (638/300 = 2.13).

Results were in-line with the huge discrepancy in mathematics instruction levels received by the groups even at the level of specific curriculum content. Regular scholastic activities included the comparison of numbers of objects and 'the implementation of counting routine' (p.23.). Indeed, counting improved in a similar extent at a statistically significant level in both groups. (Note that this also means that the extra 338 minutes of NR instruction did not improve counting beyond regular instruction.) In contrast, the groups differed from each other in number line performance and in calculation where the training group *did* but the control group *did not* receive instruction. As the authors note 'in the advanced levels of the [NR] game children had to solve summation and subtraction problems' (p.27.). So, while the NR group received fairly advanced instruction, for 5 year-olds, the control group was drawing. This instructional discrepancy is well reflected by the data: calculation performance improved a lot in the training group while the performance of the control group stayed level. This is not surprising because the control group did not receive any extra targeted instruction on calculation.

The above makes it clear that the study design is biased towards showing *any* improvements caused by NR. However, rather than any specific effects of NR the study design is merely able to demonstrate the trivial finding that if we train a group on some specific task, that group will likely improve more than another group which we do not train on that task. With regard to this, it is interesting to observe that unlike in other NR studies, non-symbolic comparison was not reported to be determined by pre- and post-tests, whereas it would have been interesting to see this measure as even the control group received (regular) instruction on number comparison. It is also worth noting that the most crucial training vs. control comparison had extreme low power (see **Table 2**) because only 9 vs. 9 children were compared, for some reason. In addition, from the very wide confidence intervals it is obvious that there was high individual variability (e.g. 52.5% and 39.4% interval width for the control and training group calculation post-tests, respectively).

The paper reaches an unjustified conclusion: 'The present RCT demonstrated the efficacy of NR for enhancing numerical skills in preschool children..'; 'NR is an effective and versatile tool for enhancing both basic and advanced numerical skills in a wide range of children' (p.27). The conclusions are unjustified because the referent of demonstrating efficacy is inadequate: The referent was practically zero level of training, a kind of activity (drawing) which cannot a priori be expected to improve mathematics skills. In other words, putative improvements were tested by essentially comparing something (NR training) to nothing (unstructured drawing activity as 'training'). Would such a comparison really justify the *specific* use of NR to train children on mathematics? Obviously not. Of course, such specific usefulness is not claimed in the paper but it is hard to imagine that this is not *implied* in a paper published on NR training with no other plausible training program included in comparisons. Inadequate designs will be discussed further in **Section 5**.

3. Focused ANS training studies

While the above studies used the NR software for fairly 'fuzzy training', Brannon and colleagues ran three studies to determine the outcome of much better controlled (more specific) ANS training. We discuss these studies in this section.

3.1 Dewind and Brannon (2012)

Dewind and Brannon (2012) addressed three questions: 1) can ANS precision be improved through training? 2) Does ANS training also improve the discrimination of other magnitudes as predicted by the so-called ATOM theory, which assumes that time, space and number are all coded/processed by the same mental representation? (Walsh, 2003)? 3) Is ANS acuity related to self-reported math performance? Twenty young adults completed 6 training sessions within 2 weeks. In session 1 they had a non-symbolic number comparison and a line-length comparison task. In each of sessions 2-5 they had 648 trials of the number comparison task and received trial-by-trial feedback. In session 6, number and line-length comparison was tested again and participants self-reported their SAT (Educational Testing Service, 2016; Zwick & Sklar, 2005) and Graduate Record Exam (Educational Testing Service, 2016; Kuncel, Hezlett, & Ones, 2001) scores.

First, a couple of words are necessary about an often used measure of ANS precision, the so-called 'w'. W is often perceived as some privileged measure characterizing the precision of the internal number representation. However, it is important to see that w simply characterizes the shape of the accuracy data arranged in a specific way across various comparison ratio conditions (see **Figure 1** and Szűcs et al. 2013 for detailed analysis). The computation of w assumes that the ANS model is valid and w characterizes the pattern of accuracy data according to the ANS model. Hence, w is entirely dependent on the accuracy data and it simply expresses the overall pattern of the accuracy data. The higher is accuracy the smaller is w and the lower is accuracy the larger is w. Hence, w is simply an alternative, model-based measure of accuracy.

@ Figure 1 about here

DeWind and Brannon (2012) posed their main question as 'the malleability of the Weber fraction in response to extended training' (p6). However, as noted above, w is a derived measure depending entirely on accuracy. So, if we train people on an ANS task and the training increases (improves) their accuracy that will inevitably lower (improve) their w. Conversely, a lower w always means higher accuracy. Hence, the question of DeWind and Brannon (2012) can be restated in a more straightforward manner as 'the malleability of accuracy in response to extended training'. Or, in an even more straightforward way as: 'Is accuracy improving on the trained task?' Likely yes, usually we would not be very surprised by such a finding. Indeed, DeWind and Brannon (2012) found that w decreased due to training. In other words this finding can be sumarized as: 'If we train people on an ANS task their accuracy will improve on the trained task.'

On another note, DeWind and Brannon (2012) conclude that there was a negative correlation between w and symbolic math test outcomes (r^2=0.28; r=-0.53) but not between w and verbal test scores (r^2=0.08; r=-0.28). However, this correlation was not robust, it dropped to practically zero when some outliers were excluded by the authors (p6; left bottom). In addition, from Fig. #7A. it seems that the significant correlation was entirely driven by a single outlier in the top left corner of the figure who had an especially large w (w=0.757; see Fig. #7A). In fact, w of this size is associated with close to chance task performance (Szűcs et al. 2013). Our impression about the correlation was confirmed when we reanalysed the original data for Fig. #7A. (We express our gratitude for the generosity of Nicholas DeWind who supplied the data). When this single outlier was removed, the w vs. math score correlation dropped to a lower level than the above noted correlation between w and verbal test scores (r=-0.21; r^2=0.044; Bca bootstrap 95% confidence interval with the single outlier removed (100,000 permutations); r= [-0.580;+0.179]. The very wide confidence interval for the original full data dataset also signals the highly unstable nature of the w vs. math score correlation; r= [-0.90; -0.04]. So, it seems that the reported correlation entirely depended on a single outlier with large w. However, because high w means poor fit to the ANS model we can also say that actually the whole correlation was driven by a participant who did not really fit the ANS model under investigation and had very low task performance. The clear instability of the effect is in sharp contrast with the conclusion of the paper which suggests that 'even in our relatively small sample of 20 subjects, acuity of the ANS was positively correlated with standardized tests of mathematical but not verbal proficiency.' As we have just demonstrated, this claim does not seem tenable even in light of analyses included in the paper which laudably reported that the correlations did not survive removing three outlier sessions.

Another point worth noting is that the study computed w based on the original ANS model which is highly imperfect because it does not factor in the impact of visual confounding parameters (see Szűcs, Nobes, Devine, Gabriel, and Gebuis, 2013; for detailed analysis of confounds; see DeWind and Brannon, 2015 for demonstrating the imperfect nature of the original ANS model; also see Fuhs et al. 2013; Gebuis and Gevers, 2011; Gebuis and Reynvoet, 2012). Hence, any (very mild) correlations between w and math scores can also be attributed to the impact of exposure and/or coping with confounding parameters: w is larger in participants who are more sensitive to visual confounds because their accuracy is typically lower. So, from this point of view it is also unclear what w exactly measured in this experiment.

3.2 Park and Brannon (2013) and Park and Brannon (2014)

Park and Brannon (2013; 2014) ran 3 training studies to determine whether an approximate arithmetic (AA) addition and subtraction task transfers to multi-digit symbolic arithmetic in adults. First, participants had a pre-test on multi-digit addition and subtraction tasks, then underwent training and then had a post-test. In the AA task participants saw animations of two dots arrays with 9 to 36 dots in each array. In one trial type participants had to decide whether the sum or the difference of the two arrays was more or less than the number of dots in a third, novel, array. In another type of trials participants decided whether the sum or the difference of the two original arrays matched the number of dots in one of two novel arrays. Trial types were mixed and they merely served to assure that participants do not develop task-specific strategies unrelated to approximate arithmetic.

Park and Brannon (2013; Exp. 1; n=52 adults) had an AA group and an unseen control group. This design cannot deliver clear data due to possible Hawthorne effects (Parsons, 1974) and due to the fact that the performance of the AA group was not contrasted with any meaningful alternative training. Exp. 2 in Park and Brannon (2013; n=46 adults) remedied these problems and had 1) an AA group, 2) a numeral ordering training group (trained to arrange triads of numbers in order) and 3) a general world knowledge training group. Only the AA group showed post-training improvement on symbolic arithmetic. Post-hoc contrasts were tested by multiple-testing uncorrected t tests.

Park and Brannon (2014; Exp 1.; n=88 adults) had 4 training groups: 1) an AA group (as in their previous study; n=18); 2) a group trained on approximate non-symbolic dot comparison (choosing the more numerous one out of two dot patterns; n=18) which is thought to improve the precision of the ANS; 3) a group trained on a Corsi-blocks type visuo-spatial short-term memory task (n=18); and 4) a group trained on number symbol ordering (n=17; as in Exp. 2 of Park and Brannon, 2013). Park and Brannon (2014) also ran an Exp. 1B with 17 participants in addition to the above 71 participants. In this experiment the appearance of the training task closely matched that of the AA condition of Exp. 1. but without the addition/subtraction requirements. Instead, participants were trained to compare and match the numerosity of dot patterns. From all the above conditions only the AA task of Exp. 1. improved post-training symbolic arithmetic performance. Ultimately, Park and Brannon (2014) concluded that training on the *manipulation of non-symbolic quantity information* led to improvements in symbolic arithmetic and hence, such manipulation may be a worthy training method for young children (p.199; however, note that their participants were adults; they connect their adult data with the controversial study of Hyde et al. 2014, discussed below).

From the critical perspective, first note that the training method used by Park and Brannon (2014) has been criticized before and it was suggested that it produced data similar to those we could expect from a non-learning observer (Lindskog and Winman, 2014). Besides this there are still two major flaws in the conclusions. First, it is clear that only the AA task led to transfer to symbolic arithmetic. However, from the design and the results it does not follow at all that the non-symbolic nature of the AA task is a *necessary*[1] component of successful transfer. Second, from Park and Brannon (2014; Exp. 1 and Exp. 1B) it is also clear that a non-symbolic arithmetic comparison task on its own is in fact *insufficient* to generate transfer to symbolic addition and subtraction tasks. However, the non-symbolic comparison task serves as the most important measure of the precision of the ANS.

[1] Note the difference between sufficient and necessary: A sufficient condition is one which is enough to lead to an outcome *on its own*. A necessary condition may not be enough to lead to an outcome on its own (ie. it may not be sufficient) but it must be one condition to fulfil perhaps together with other conditions to achieve the outcome.

Hence, it seems that training the ANS on its own was not able to improve arithmetic performance at all.

Let's evaluate the above statements in detail. First, is improving the ANS a *sufficient* condition to have transfer? Put otherwise, if we *solely* sharpen the *precision* of the ANS without any other training will that improve symbolic addition/subtraction performance? A strong interpretation of the ANS theory would surely predict this and this is in fact implied in many papers which claim to have demonstrated correlations between the ANS and symbolic arithmetic (see Szűcs et al. 2013 for a critical review). Park and Brannon tested this question in 2 experiments (Park and Brannon, 2014; Exp. 1 and Exp. 1B). Both experiments returned negative findings on more than one level. First, ANS training operationalized as non-symbolic dot comparison training did not lead to transfer in any of the experiments. Second, the effect size of the correlation between w and addition/subtraction performance was practically zero ($r=-0.07$; $r^2=0.005$; $p=0.509$; see top left panel in Fig. #7 of Park and Brannon, 2014) also adding to numerous negative findings (see Szűcs et al. [2014] for review). To put it clearly: w was unrelated to math outcomes and ANS training did not improve symbolic outcome measures. So, ANS training is not a sufficient condition to have transfer. This conclusion poses a serious challenge to claims that ANS is (causally) related to math performance.

Further, is ANS training a *necessary* condition to have transfer? That is, must ANS training be a component of a successful training programme? The interpretation of Park and Brannon (2014) is ambiguous and reflects a strong bias towards the ANS theory. First they say: 'the more active process of manipulation of mental representations is the critical mechanism underlying the observed transfer effect' (p198.). In fact, the authors seem to be conscious of this option already in Park and Brannon (2013; p6) which says: 'Another possibility… is that the training and transfer effects in the current study reflect facilitations in cognitive processes related to addition and subtraction'. However, testing this option was not built into the design of Park and Brannon (2014) whereas it would have been fairly straightforward (see later). Clearly, the results do not provide any evidence that a key element of the studies was training the ANS in any way. As noted above, ANS training on its own did not lead to transfer and w was unrelated to math. So, the authors themselves note (see the preceding quote) that the most likely explanation for the data is that training on addition and subtraction per se led to transfer. However, Park and Brannon (2014) finally conclude: 'our study demonstrates that providing … multiple sessions of approximate arithmetic training improves exact symbolic arithmetic' (p199.). So, while they recognize that the manipulations in the task were key, they then blur this interpretation together with the fact that the stimulus material happened to be dot patterns. Note that their statement can be accepted to be literally true: their training *was* approximate arithmetic training. What causes the problem is the *implied necessary* nature of the *ANS element* of this training whereas no data supports this implication. It is much more likely that the *crucial* element of the training was *practicing addition and subtraction irrespective* of the ANS element. A straightforward explanation for the authors' explanation is bias towards interpreting the outcomes from the view of the ANS theory rather than considering alternatives. We suggest that a simple summary explanation for the findings is that the non-symbolic nature of the AA task was irrelevant, what mattered was the *operations* trained. In other words, the AA task led to transfer because it implicitly trained addition and subtraction and not because it had any connection with the ANS. This conclusion is line with the results from the non-symbolic number comparison training tasks of Park and Brannon (2014).

An additional result from Park and Brannon (2014) supporting our above conclusion is that w, the most important marker of the ANS (see above on the nature of w), did not improve in response to AA and number comparison training ($F(3,62)=1.352$; $p=0.266$; p.195; right bottom). The authors' presentation of these statistics reflects strong bias for the ANS theory. After they communicate the above non-significant ANOVA outcome they go on and collapse the AA and the number comparison groups (note that such 'flexible analysis' is likely to generate false positives; see e.g. Simmons et al. 2011) and the other two groups and then compare the former to the latter two groups with a most probably ad-hoc uncorrected t test ($t(59)=1.984$; $p=0.052$). They then interpret this test outcome calling it a 'strong trend' and also put emphasis on it in the Discussion (p197) saying that 'numerical comparison training showed some evidence of improved w … suggesting a near transfer effect'. It is clearly an overstretch to interpret the outcome of a non-significant F test and a most probably unplanned non-significant t-test in such clear terms[2]. This over-interpretation suggests strong bias to support the ANS theory.

Park and Brannon (2014) ran their Exp. 2. to exclude the possibility that participants used verbal addition/subtraction strategies in their non-symbolic training task. In this experiment they show that verbal interference decreased exact symbolic arithmetic performance but not non-symbolic arithmetic performance. So, they conclude that the non-symbolic training task engaged non-verbal processes. While the literal interpretation of this conclusion is fine, it is important to realize that a 'nonverbal' training element does not automatically mean that the crucial non-verbal element has any relation to the ANS. Making such a connection is another unjustified implication. In fact, as other results show, the ANS training was ineffective. So, the 'nonverbal' nature of the task may mean that, for example, visuo-spatial or attention processes rather than the ANS was improved by the task. For example, we found that ANS task performance was related to sustained attention rather than to numeracy (Szűcs et al. 2014). So it follows that Exp. 2. does not affect the critical points noted above regarding the interpretation of the data. Also, it is important to notice that there was strong asymmetry between the stimulus material used in the symbolic and non-symbolic conditions in Exp. 2. While hundreds of symbolic arithmetic problems were used, only three log differences were used in the non-symbolic task. In addition (or, perhaps for this reason), the symbolic task was more difficult reflected in much longer reaction times than in the non-symbolic task (6.57 and 6.09 seconds in the symbolic task vs. 0.879 and 0.913 seconds in the non-symbolic task). Such large task difficulty discrepancy can easily impact on the data. It may make much more sense to revert to verbal strategies in the more difficult task.

4. Brief exposure to ANS tasks which do not qualify as training studies

Two further studies are important to discuss even if they are *not* proper training studies because they make very strong claims about the usefulness of ANS exposure for mathematical improvement in children (Hyde et al. 2014; Wang et al. 2016). These studies provided a *brief single* session exposure to ANS tasks and concluded that this exposure improved symbolic math performance right afterwards. Results are clearly overinterpreted in both papers suggesting that they found it a 'fact that a single session of practice on an approximate number task can improve' symbolic math performance (Hyde et al. 2014; p105) and that 'there is a causal link from ANS precision to symbolic math performance' (Wang et al. 2016; p95). These overinterpretations from brief single session data with a few practice trials (72 training trials with 8 practice trials in Hyde et al. 2014; 30 training trials with 4 practice trials in Wang et al. 2016) are even more surprising in the context of the many inconclusive and negative results from more appropriate training studies published before and discussed above.

4.1 Hyde et al. (2014)

Hyde et al. (2014; Exp. 1.) trained 4 groups of 24 grade one children (96 in total). Groups received one of four kinds of training: 1) non-symbolic number addition, 2) line-length addition, 3) non-symbolic number comparison and 4) brightness comparison. Children had 50 training trials followed by 10 easy and 10 moderately easy symbolic arithmetic test items. Then children had 10 more training trials and 20 moderately difficult symbolic test items. In the non-symbolic addition task children saw dots for 1 second, dots moved out of view in half a second, there was a pause for half a second and then another dot pattern for 1 second. After this, children decided whether a third dot pattern had more or less dots than the sum of the two previous dot patterns.

The question concerning whether the 4 training tasks improved subsequent symbolic math performance was evaluated by two F tests and subsequent multiple testing uncorrected two-tailed t-tests. A simple Bonferroni correction for the 6 relevant comparisons per F test (4 × 4 table with 6 unique pairwise comparisons) would require $\alpha=0.05/6=0.0083$. None of the reported t-tests reached

[2] Note that the degrees of freedoms for the t test does not seem to fit as the collapsed groups should have had n=18+18=36 and n=18+17=35. So we could expect degrees of freedom of 36+35-2=69. However, most probably t(69) was mistyped as t(59) in which case t(69)=1.984 should be associated with p=0.0512 (reported as t(59)=1.984).

this significance level (uncorrected p value range for significant tests: 0.0133 – 0.0482). Hence, the robustness of results is dubious as most comparisons are likely to be n.s. even if less conservative methods than the Bonferroni correction were used.

The authors reported that symbolic math performance was faster and less error prone after both kinds of ANS training than after line addition and brightness comparison training (Fig. #4). The authors suggested that these results 'provide evidence that the ANS plays a functional role in symbolic arithmetic' (p99.). However, a very simple alternative explanation is that the numerical (ANS) tasks simply primed attention to numerical information and activated general number knowledge related to addition and comparison while this was not the case for the other training tasks which did not share any component with the test task. That is, the results provide absolutely no evidence that the ANS is functionally related to symbolic math (the paper implies that this relation in inherent). However, it is strongly implied that the presence of the ANS element is a *necessary* cause of the observed improvement.

An unlikely explanation is given about why the results cannot simply reflect practice with addition and comparison processes. First, similarly to DeWind and Brannon (2012), the authors argue that the data is in *disagreement* with the strong version of a generalized magnitude system posed by the ATOM (Walsh 2003) theory (p100, top) because the number line addition and the brightness comparison task did not improve symbolic math performance. So, they argue that the physical magnitude system is *distinct* from the number magnitude system. Right after this, the next argument is that the data cannot reflect general practice with number comparison/addition processes because there was no symbolic math improvement after line summing and brightness comparison training whereas these tasks involved 'the same cognitive operations (ordering, comparison and/or addition)' (p100, top) as the numerical conditions. However, right before this argument, the authors had just concluded that the physical magnitude system is *not* overlapping with the number magnitude system. So, there is absolutely no reason to assume that the 'addition' and comparison processes operating on these representations are 'the same'. However, if they are not the same then the unique training of number addition and number comparison processes (rather than ANS) training can still be contributing to improved symbolic math performance. In summary, the authors first discard the ATOM theory (Walsh 2003) and then they use an assumption of the discarded theory to justify their next argument.

Hyde et al. (2014; Exp. 2.) went on to test whether any performance improvement after brief ANS exposure was specific to mathematics. They used exposure to 1) non-symbolic numerical addition and 2) non-symbolic brightness comparison in Exp. 2. To this end the post-exposure performance on a symbolic math test and on a sentence comparison test was compared across the two conditions. Unsurprisingly, it was found that only symbolic addition performance but not sentence comparison performance improved after the non-symbolic addition exposure. Performance did not change after exposure to the brightness task. Note that this outcome can also be explained by the above two alternative explanations: *attention* was directed to number in the non-symbolic addition training but not in the brightness training and *addition* was trained in the ANS addition task but not in the other task. So, there are at least two reasons for the pattern of results which have nothing to do with the ANS element of the exposure. Whereas these alternative explanations were clear even after Exp. 1., the first one was never considered and the second one was discarded based on an inconsistent argument (described above). So, rather than testing any of the above very likely alternative hypotheses, Exp. 2. tested a fairly unlikely null hypothesis based on an already discarded theory. This is exactly the design approach criticised by Meehl (1967) in his classical article (see more on this later). The uncertain results are highly over-interpreted, the paper concluding that there exists 'a causal relationship between non-symbolic approximate number and exact, symbolic arithmetic by children.' (p105).

4.2 Wang et al. (2016)

Wang et al. (2016) claim to demonstrate that 'temporary modulation of ANS precision changes symbolic math performance'. First, this claim seems somewhat of an oxymoron: if ANS is a relatively stable representation in the mind how can we 'temporarily' modulate it? At least, the same research group's previous papers suggest that the ANS is a robust enough representation so that we can base mathematical disability diagnoses on its status. So, would it not be much more likely that, rather than temporarily improving a supposedly stable representation we can rather improve access to it perhaps by directing attention to it?

In the study the authors replicated the non-symbolic numerical comparison condition of Hyde et al. (2014) using only 30 training trials with 40 five year 4-month-old participants. 20 children proceeded from easy to hard ANS comparisons (easy-first group) while 20 children proceeded from hard to easy comparisons (hard-first group). Half of the children in each group had a symbolic math test after training while the other half had a vocabulary test. This design step is hard to justify as it deprived the researchers from potentially important within-subject data. All 20 children should have had both tests in counterbalanced order, this is well possible at the age group tested. A second crucial design problem is that there was no pre-exposure symbolic math and vocabulary test to measure baseline performance. Hence, it cannot be determined whether children in different groups had very different symbolic math levels to start with. This omission is makes it questionable whether the results of the experiment can be interpreted at all. Third, it is to note that the study did not seem to correct for multiple comparisons.

After exposure, children in the easy-first group showed much higher symbolic math performance than children in the hard-first condition (percent correct: 82.78% vs 60.56%). Because there was no pre-exposure test, we cannot conclude about any within-group performance change. However, the post-exposure symbolic math performance in the hard-first condition was even worse than the post-exposure performance of the other children in the vocabulary test (67.50% and 67.91%). The very low performance of the hard-first group on the symbolic math test (if it is not attributable to a large pre-exposure between-group difference) may mean that the children 1) had no idea what they had to do because the task was initially so difficult and/or 2) their performance did not improve because they did not pay attention to number due to the initially large task difficulty.

In fact, we argue that the observed very low performance is incompatible with any ANS activation account because the ANS is supposed to be activated by the mere presentation of non-symbolic numerals. Rather, it seems that because children found the initial discriminations too difficult, they were guessing in many trials which is reflected in their extremely low accuracy rate: 60.56% (first paragraph in p89). Regarding this, it is important that surprisingly, this is one of the few papers where the authors do not use w to characterize number comparison performance; however, it is possible to estimate it from the accuracy data. In our previous investigation (Szűcs et al. 2013) we tested twenty 7-year 5-month-old children in an approximate number comparison task and found an accuracy level of 62.5% which corresponded to a w value of 0.77. Hence, we can expect an even larger w value for the children tested by Wang et al. (2016; lower accuracy means a larger w value). However, even a w value of 0.77 is already much higher than $w \approx 0.4$ which was thought to characterize dyscalculia by the same group of authors (Mazocco, Feigenson, & Halberda, 2011). So, based on the authors' previous papers, we could assume that perhaps all the children in the hard-first condition had very severe dyscalculia. Henceforth, it would not be surprising that their performance did not improve. Alternatively, it is much more likely that the children were confused about the task, they were not doing it properly in many trials and thus that they ended up with very low accuracy. This reasonable assumption would of course disqualify all findings because it would mean that in one of the conditions children were not doing the intended task properly. Further, at this low accuracy level individual variability is also of great concern: in a study of 7-year-olds and adults we found large individual differences in non-symbolic number comparison performance (Szűcs et al., 2013). This has not been considered here, either.

The authors stated that they chose a group size of 10 based on a power analysis which assured 80% power taking into account the results of Hyde et al. (2014). However, in their Exp. 2 Hyde et al. (2014) reported $t(46)=2.814$ with 24 participants in each of two groups. This translates into an effect size of $D=2\times t/\mathrm{sqrt}(48) = 0.8123$ (Fritz et al. 2012; see derivations in Szucs and Ioannidis 2016). First, such high effect size is clearly inflated: it is well known that small scale studies vastly overestimate effect sizes (Schmidt, 1992; Button et al. 013). However, even if we consider this inflated effect size of 0.81 and that Wang et al. (2016) only had 10 participants in a group (which is small by usual standards), then we can compute power = 0.4051 for an independent sample t-test ($\alpha=0.05$). Such pairwise tests would have been necessary to determine whether important contrasts were statistically significant (note that they were never reported numerically just in the form of asterisks in Fig. #2.). Further, if we compute power for a fixed effects ANOVA with df = 1,36 for a large effect size ($f=0.4$

in GPower which is about equivalent to D=0.8; see Cohen (1988)) then we still only get power = 0.692. However, as noted, it is well known that published studies overestimate effect sizes (Schmidt, 1992). So, it is more realistic to compute power for small, medium and large effect sizes rather than for an effect size of 0.81 (Sedlmeyer and Gigerenzer (1989; Szucs and Ioannidis, 2016). For these effect sizes the power of the independent t-test ranged between 0.1-0.4 (**Table 2)**. Power for similar small and medium effect sizes for F tests (df=1,36) are power = 0.152 (f=0.15; D=0.3); power=0.455 (f=0.3; D=0.6). Hence, the power of the study was much lower than declared.

The main conclusion of the study, similarly to Hyde et al. (2014) is that brief exposure to an ANS number comparison task improves symbolic math performance. However, the data of both Hyde et al. (2014) and Wang et al (2016) can be explained by the same alternative explanations: Both studies may well just have primed attention to numerical information. Hence, we suggest that a succinct summary of the most likely explanation of the findings of Hyde et al. (2014) and Wang et al. (2016) is: If we direct attention to number that will boost performance on tasks involving number but not on other tasks.

A note is that the low number of trials and training trials is clearly a problem in the brief ANS exposure experiments. It is a trivial fact that initial task performance improves quickly in nearly anything we can test. This is exactly the treason that good quality experiments have many training trials if this is possible. For example, if brief ANS exposure experiments simply measure the impact of directing attention to numerical information in general than such attentional effects can be expected to be particularly strong in the beginning of experiments, especially with few training trials.

5. Some general points

Below we highlight some major problems which recur in studies.

5.1 Low power, high false report probability, exaggerated effect sizes

Table 2 shows power to detect small, medium and large effect sizes as defined by Sedlmeyer and Gigerenzer (1989; Power calculation parameters are presented in the caption of Table 2. For a detailed exposition on power, effect size and false report probability see Szucs and Ioannidis, 2016). Only Wilson et al. (2009) and Obersteiner et al. (2013) had power > 0.5 to show medium sized effects (power range: 0.17 – 0.69) and studies had very low power to show small effects (power range: 0.1 – 0.3). The consequence of low power is not only that real effects may be missed but also very high false report probability and exaggeration of effect sizes measured in studies (Szűcs and Ioannidis 2016; Button et al. 2013).

It is important to point to two frequent misconceptions: First, it is often thought that a large (statistically significant) effect size in a study with low power means that a finding can be particularly trusted because 'if even a small study could detect an effect it must be really robust'. However, (perhaps counterintuitively) low power is inevitably associated with large effect sizes because with low degrees of freedom only large deviations from the value associated with the null hypothesis can reach statistical significance. The key test of these detected effects is not whether they look large in a single study but whether they are replicable. Second, it is often thought that if a study has detected a statistically significant finding then that finding must be accepted as a 'fact', or that at least that particular finding is highly robust even if it comes from an underpowered study. These are wrong assumptions: Any findings from underpowered studies have high false report probability irrespective of whether the findings are statistically significant or not (Button et al. 2013; see detailed modelling in Szucs and Ioannidis, 2016). In fact, usual power limitations in psychology and neuroscience mean that most publications report exaggerated effect sizes and have high false report probability (Button et al. 2013).

Overall, low power can result in widely varying statistically significant findings as one underpowered study may find a large effect size into one direction while another underpowered study may just find a large effect size into the opposite direction. The lesser is power across studies the more variable findings will become. These problems are of particular concern regarding the very small scale study of Wilson et al. (2006b) which has nevertheless been cited 50 times with 25 citations claiming that training improved arithmetic (**Table 3**; see analysis later). Such citations are clearly unjustified not only in light of the very low power of the study but also because it did not have a control group. Most other studies discussed here also had modest power to detect small and medium effects (**Table 2)** In general, especially in light of the current replication crisis of psychology and neuroscience (Nosek et al. 2015) it is extremely important to interpret findings from low powered and inconclusive studies very cautiously (Button et al. 2013).

5.2 Endemic lack of multiple testing correction

Surprisingly, with the sole exception of Obersteiner et al. (2013), who used multiple-testing corrected Scheffe tests, none of the studies noted that they used any multiple testing correction for pairwise comparisons following ANOVAs. In fact, they explicitly seem to suggest that they relied on simple t-tests in pairwise comparisons. One study even lacked any clear reporting of pairwise comparisons relying only on an asterisks notation (Wang et al. 2016). The lack of multiple testing correction is further exacerbated by the fact that sometimes non-significant ANOVA outcomes were followed up by such t-tests (Park and Brannon, 2014) and/or sometimes marginally non-significant, uncorrected t-tests were treated as statistically significant outcomes and interpreted in discussions (Wilson et al. 2006b; Wilson et al. 2009; Park and Brannon, 2014). However, multiple testing correction is necessary when qualifying several pairwise comparisons from ANOVAs and uncorrected tests should not be interpreted. Regular reliance on the above mistaken statistical inferential approaches can largely inflate the number of false positive findings.

5.3 The use of ANCOVA

It is invalid to use ANCOVA to ‘correct for’ pre-study group differences. Put otherwise, ANCOVA cannot be used with a covariate which is significantly different along the grouping variable(s) of interest (see Miller and Chapman, 2001; Porter and Raudenbush, 1987; Evans and Anastasio, 1968; Lord, 1969; Lord 1967). Such use of ANCOVA can substantially distort the data, render grouping variables meaningless and can result in entirely spurious statistically significant analysis outcomes. Yet, ANCOVA is frequently used in this incorrect manner, exactly with the intention of treating pronounced and perhaps statistically significant pre-intervention mathematical score group differences as non-significant (Sella et al. 2016; Park and Brannon, 2013; Park and Brannnon, 2014; Obersteiner et al. 2013). For example, Sella et al. (2016) based their whole analysis on ANCOVA even when they seemed conscious of the above problem because Footnote #3 (p.23.) communicates that pre-test scores did not differ between training and control groups. However, this contradicts the authors’ justification of using ANCOVA on the same page where they noted that they used ANCOVA because the two groups ‘substantially differed before training’ (p.23. bottom right; and this is indeed the case for most variables by looking at their Table #1).

In general, if experimental and control groups are substantially different from each other on some pre-test variable then there is no method which could achieve that we would be sure to know how the groups would perform were they *not* different from each other (Miller and Chapman, 2001; Porter and Raudenbush, 1987). First, a strong pre-study difference may be avoided by proper individual randomization of training assignments and having large sample sizes which allow for more adequate randomization. So, besides low power the expected lack of adequate randomization and consequent large pre-study group differences is another problem of small scale intervention studies. Second, if pre-study training group differences exist then they simply cannot be ‘corrected for’. Rather, differences along important variables must explicitly be factored into analyses, for example through regression models (Miller and Chapman, 2001). The predictive value of these pre-study differences must then be communicated rather than just noting that ‘they were controlled for’ as lack of appropriate detail renders analyses meaningless. In addition, calculating effect sizes and confidence intervals for between/within group differences can also be very informative (see the analyses of Räsänen et al. 2009 and Obersteiner et al. 2013 and the additions to them in this paper). Third, if large pre-study group differences are unavoidable then the study should be replicated with a different pre-study pattern of training and control group participant assignment before any conclusions can be drawn. This is especially so if researchers wish to make strong statements like for example, ‘we found it a fact’.

5.4 Design: Good and bad choice of control activities and time on tasks

Three NR studies have clearly inadequate design: Wilson et al. (2006b) did not have a control group at all, Wilson et al. (2009) contrasted NR training with reading training and Sella et al. (2016) contrasted NR training with unstructured drawing activity. In contrast, Räsänen et al. (2009) and Obersteiner et al. (2013) had much more meaningful designs and contrasted alternative forms of math trainings rather than math training and non-math training or nothing. Notably, Räsänen et al. (2009) found that GraphoGame-Math was superior to NR and Obersteiner et al. (2013) found no difference between ANS based approximate and exact numerical training.

The important general conclusion to draw is that it does not make sense to contrast target-related interventions with completely target-irrelevant ones (as in Wilson et al. 2009; Sella et al. 2016). Such designs may also largely exaggerate group differences in the amount of mathematics instruction received by groups adding another confounding factor as in Sella et al. (2016; 638 vs. 300 minutes). In other studies the problem is exacerbated by deliberately adding more mathematics instruction than control instruction as in Wilson et al. (2009; 4:6 reading vs. mathematics training ratio). Such designs strongly bias studies to detect larger effects of one than the other intervention.

Overall, if our objective is to claim that our intervention is especially useful for improving mathematics than it does not make much sense to contrast our math-related intervention with some non-math related intervention. For analogy, if we want to claim that our method is especially useful for teaching children to swim than it is not fair to first test children on swimming and then contrast our swimming training method with a running, or casual walking training method. Could we declare surprise (significance) when we find that all children who took our (perhaps amateurish) swimming class can at least stay afloat just coughing up a bit of water but all the children from the running group sank? Could we then market our swimming training method to parents as a worthy method to try with their children? Rather, a fair comparison is to contrast alternative proposed interventions (as in Räsänen et al. 2009; Obersteiner et al. 2013) and/or to contrast our favoured intervention with an already established and well-working target-related intervention (e.g. traditional math-education) with our favourite intervention. After all, why bother with introducing our new method if another, already established method works perfectly well, and perhaps much better than our method?

Overall, any meaningful qualification of a training programme should compare the programme to another training programme which is either used, or can be expected to be used. Such correct designs were chosen by Räsänen et al. (2009) and Obersteiner et al. (2013). Overall, the real question is not whether we should use a training programme which is perhaps providing marginally better improvements than zero but rather, *which* successful targeted training programme should we use?

5.5 Not following up the most important alternative hypotheses

Choosing suboptimal alternative trainings is also a problem in studies with more focussed ANS tasks. As we discussed, while Park and Brannon (2013) did consider that perhaps general practice with addition and subtraction processes rather than ANS experience explains their data they chose not to test this very likely alternative hypothesis in Park and Brannon (2014). Similarly, Hyde et al. (2014) discarded the ATOM theory in Exp. 1. but then they still based their null hypothesis in Exp. 2 on the predictions of an already discarded theory rather than testing two very likely alternative hypotheses.

The above examples seem similar to those discussed by Meehl (1967). Meehl suggested that studies strongly biased towards some theoretical explanations typically choose to test very unlikely null hypotheses rather than contrasting their favourite ideas with more likely alternative hypotheses. He also directed attention to the 'liberal use of ad-hoc explanations', 'use of complex and rather dubious auxiliary assumptions which are required to mediate the original prediction and are therefore readily available as (genuinely) plausible "outs" when the prediction fails'. Such problems are very evident in Wilson et al. (2009) who explained the lack of an expected finding by reverting to tautologic and circular ad hoc explanations about their participants assumed unmeasured internal characteristics and selective citing of positive evidence only (se discussion above). Again, considering the current replication crisis of psychology it is more important than ever to refrain from such unjustified ad hoc arguments and to test the most important alternative hypotheses rather than less likely and less important ones.

6. Citation bias in the ANS training literature

Table 3. summarizes citations to the papers discussed (see Methods in **Appendix 2 and the collection of citations and references to citing articles in Supplementary Material; date of study: May 2016**). We identified 85 citing articles making 285 citations to the papers discussed here.

Strikingly, in contrast to the serious problems with most papers discussed above only 13 citations from 9 papers made a critical comment about the papers and/or noted the lack of training effects in at least one of the ANS intervention studies (Chen, Q. & Li, J., 2014; Jang, S., Cho, S., 2016; LeFevre, J., 2016 ; Leibovich, T. & Ansari, D., 2016; 5:Lindskog, M., Winman, A. & Juslin, P., 2013; Lindskog, M. & Winman, A., 2016; Räsänena, P., Salminena, J., Wilson, A., Aunioa, P., & Dehaene, D., 2009; Salminen, J., Koponen, T., Leskinen, M., Poikkeus, A., Aro, M., 2015; Torbeyns, J., Gilmore, C. & Verschaffel, L., 2015). * OBERSTEINER REMOVED

Nine citations (a subset of the 13 citations mentioned above) from 6 papers offered more specific critical comments. Two of these critical comments regarded the lack of control group in Wilson et al. (2006b; Lindskog et al. 2013; Räsänen et al., 2010). Torbeyns et al. (2015) notes that most of the ANS intervention studies have serious methodological flaws, specifically citing not having proper control groups (p 106). Another paper suggests that the arithmetic training imbedded in the approximate arithmetic task of Park and Brannon (2014) rather than the ANS acuity training is likely responsible for reported improvements (Leibovich & Ansari, 2016). One paper (described above) was entirely devoted to the critique of Park and Brannon (2014; Lindskog and Winman, 2016). Another paper by Lindskog et al. (2013) reported their own attempt to replicate the results of DeWind and Brannon (2012). After controlling for perceptual cues, Lindskog et al. (2013) found no effect for learning transfer from ANS to symbolic math. With regards to DeWind and Brannon (2012), Jang and Cho (2016) point out that the inconsistencies in results between this study and others with similar designs may be due to differences in the dimensions used for visual stimuli and in the visual complexity of the tasks. Without citing specifics, LeFevre (2016) reports that the studies were not 'uniformly' successful in showing transfer between ANS training and symbolic math performance.

Considering all 50 citations to Wilson et al. (2006b) we can observe that only 4% of citations identified at least one problem in the study. Moreover, considering that all discussed papers together received 253 citations, only 6% of all citations raised any problems and 4% discussed problems in more specific terms. Considering that science is supposed to progress based on challenging controversies, the lack of critical comments is highly notable because several problems can be raised with regards to most studies (except Räsänen et al. 2009 and Obersteiner et al. 2013).

Half of the 50 citations to Wilson et al. (2006) cited it claiming that NR training improves arithmetic performance and 14% of citations suggested ANS plays a causal role in arithmetic improvement. These claims are clearly unfounded in specific terms considering the small size of the study and that it did not have a control group. Nearly all papers citing Hyde et al. (2014) and Park and Brannon (2013) made the same claims. Notably, even Räsänen et al. (2009) was cited once stating the very general sounding claim that they demonstrated 'a link between training on approximate arithmetic and symbolic math ability' (Park and Brannon, 2013; p. 5). However, Räsänen et al. (2009) merely found that NR training improved speed on symbolic number comparison while training did not affect any other symbolic task. So, the overgeneralization of the citation is clearly unfounded.

Overall, 55% of the 253 citations suggested that ANS training improves arithmetic and 38% of citations suggested that ANS plays causal role in this improvement. It is also notable that while we counted 30 citations from 22 review articles in our sample only 2 such citations from 2 review articles noted any critical comments about the papers discussed here (Leibovich and Ansari, 2016; LeFevre, 2016). Considering the serious controversies we analysed before we conclude that several citing studies demonstrate a strong bias favouring the idea that the ANS is causally related to symbolic mathematics. Clearly, studies must take a more critical approach to evaluating evidence rather than just restating conclusions from highly controversial papers.

@ Table 3 about here

7. Recommendations

7.1 Design

Measured training transfer effects are the consequence of the overlap between the mental representations and processes *affected* by training and the representations and processes necessary to carry out the tasks used as outcome measures (**Fig. 2A.**). It is important that the training can have impact on representations and processes not intended to be affected, so we have to be careful when evaluating what exactly was trained and what exactly outcome measures represent. For example, a researcher may expect that NR training only sharpens ANS precision and hence, may conclude that any post-training improvement in mathematics performance is due to improved ANS precision. However, as discussed above, it is clear that NR affects many more representations and processes beyond the ANS. Or, another researcher may assume that a non-symbolic dot addition task sharpens ANS precision only whereas the key impact of the training may be general addition practice and some attention training irrespective of the non-symbolic material used. In both above cases, it is hard to decide what exactly potential transfer effects may be related to without further qualifying experiments. As we suggested, several studies seem to have avoided to address the most important questions regarding ANS training whereas it would be straightforward to set up tests. Here we recommend clear designs.

The crucial operational design question regards task and stimulus specificity. Stimulus specificity may be more related to the question of representations used to code information (e.g. symbolic or non-symbolic representation) while task-specificity may be more related to the processes run on representations (e.g. addition, subtraction, comparison). Naturally, representations and processes may interact, for example, some processes (e.g. some visual addition or subtraction algorithms) may be available for symbolic but not for non-symbolic stimuli. Systematic design focussed on stimulus and task specificity may also be able to uncover such interactions.

A simple design suggestion is given in **Fig. 2B**. Initially we assume that participants would be primary school children. Optimally, outcome measures should be taken before the study, right after the study and a longer time period after the study. It is also beneficial to take outcome measures during the study more than once to track the rate of change in outcome measures. The design (**Fig. 2B.**) considers the stimulus and task specificity of the training task along two levels. The trained stimulus material can be non-symbolic and symbolic and the trained task can be addition or comparison. The outcome measures are symbolic addition and comparison (shaded area in **Fig. 2B2**.). We hypothesize that that symbolic addition outcome will improve more in any conditions involving addition than in conditions with comparison. We also predict that the best training results will be achieved in the 'symbolic training material with addition' condition because this has the most overlap with the symbolic addition outcome task.

More complex designs could add more levels to both the trained operations (addition, comparison, subtraction, number ordering) and to outcome measures. We predict that a given outcome measure will improve when a particular operation is trained irrespective of the stimulus material. We also predict that symbolic training will provide better results than non-symbolic training due to the enhanced precision of symbolic representations.

It can be raised that the symbolic stimulus / symbolic test outcome measure option is too direct and of course the best improvement can be expected when symbolic material is used to train symbolic operations. However, in relation to school arithmetic we expect children to work with symbolic numbers and the human specific mathematics they have to learn is based on symbolic numbers. (Would we be happy if children only learn to pay an approximate sum in the shop, or get approximately home after school?). Moreover, we know from the regular school curriculum that training with symbolic operations leads to improvement in symbolic operations. So, what is the point of training symbolic operations indirectly (through the ANS, by using dot patterns) when in fact we can train them directly probably with better outcome?

Naturally, it could be argued that non-symbolic training may be better for 1) small children and 2) for people with poor mathematics. However, in that case studies should still contrast whether the proposed alternative indirect training provides better outcomes than the more direct training with symbolic numerals in certain groups and certain age ranges. E.g. it may happen that training small children with non-symbolic addition before they learn numbers is beneficial. However, will this intervention deliver any specific long-term improvement once children start symbolic learning besides an initial (perhaps irrelevant) boost? In order to test such questions, designs can be complicated by testing various age groups, including kindergarten children. In such case the crucial question would be whether using non-symbolic material has any benefit over symbolic material at an earlier age than primary school.

@ Figure 2 about here

7.2 Recommendations: reporting

Several good general recommendations for improved reporting are given by Simmons et al. (2011) and specifically for training studies recently by Moreau et al. (2016) and Green et al. (2014). We recommend *pre-study* power calculation for small, medium and large effect sizes (see e.g. Szucs and Ioannidis, 2016). We recommend pre-registering all studies before they start and publishing all raw data with the primary publication from the intervention study (obviously, a pre-requisite of using this data must be citing the publishing article). It is essential to determine and publish pre-study group differences in important variables. Relying solely on gain scores is inadequate and their use can be misleading (see for example Moreau et al. 2016). Tests should be corrected for multiple comparisons to avoid the inflation of Type I error. Rather than relying on point estimates of parameters, it is more informative to provide interval estimates such as confidence intervals. If normality is not achieved than bootstrap methods could be used. Effect sizes need to be calculated. Confidence intervals should not be confused with Bayesian credible intervals which provide more useful information than confidence intervals (Hoekstra et al. 2014).

Studies should state participant numbers clearly upfront; for example it may happen that a study states that 22 children were recruited (Wilson et al. 2006b), then goes on to say that 13 of these children were selected for the study and ultimately says that the ‘final sample’ of the study was 9 children. Such descriptions are unfortunately fairly typical in the developmental literature. However, rather than describing the process of losing participants it is much more straightforward and informative to state final participant numbers to start with and describe details afterwards.

Discussions must avoid post-hoc theorizing and unnecessarily complicated arguments perhaps taking a biased view of evidence. A related simple fact is that low power leads to highly variable results in studies. This will facilitate looking for alternative mediating explanations and starting theorizing about these (Schmidt, 1992). However, variability in findings may entirely be due to low power. So, our primary job is to increase power and report findings clearly rather than unnecessary (and often confusing) theorizing.

7.3 Overcoming the citation bias: Critical analysis is needed

Our citation analysis demonstrates the lack of critical comments regarding the studies discussed and that many claims supported by citations were unfounded. It is important that we break with the ‘business as usual’ tradition and take a more critical stance when evaluating evidence. This will not only result in more efficient use of research funds but will also speed up scientific progress.

8. We do like non-symbolic math training

It is important to caution that we are not against the use of non-symbolic information and manipulatives. We are confident that manipulatives, concrete countable objects, and other forms of learning which focus on general counting, comparing, and manipulation skills can be useful in early number training (Dyson, Jordan, & Glutting, 2013; Fennema, 1972; Suydam & Higins, 1977). In fact, they have been used for hundreds of years (Froebel, 1899; Montessori, 1882) and are still currently used with positive effect in many formal kindergartens and school systems (Carbonneau, Marley, & Selig, 2013; Sowell, E. J., 1989). As Dyson, Jordan, and Glutting (2013) show, counting, comparing and manipulating sets can help children improve their sense of number (number sense here defined as in Jordan et al., 2012, p. 2), which can lead to improved performance in the classroom. In fact, our own research also confirms that concrete three-dimensional spatial building ability is related to numerical understanding in 7-year-old children (Nath and Szűcs, 2014).

What we argue against is biased designs and interpretations and incorrect use of statistics. For example, we do not see much evidence that ANS training *specifically* improved anything in the reviewed studies. This may be because 1) ANS training is already inefficient in the age groups tested and/or 2) because dot pattern comparison and/or their mental manipulation is not very useful in general. This last statement does not exclude that other non-symbolic math training works. We need properly designed studies with balanced interpretations to determine similar questions.

9. Conclusions

Our critical analysis reveals a large number of problems in the ANS training literature. Several studies are poorly designed, lack power, use inadequate statistical procedures (e.g. illegitimate use of ANCOVA and lack of multiple testing correction) and rely on highly biased inference. We conclude that with the exception of Räsänen et al. (2009) and Obersteiner et al. (2013) all other studies discussed here had inadequate design and/or inference. The above two studies could not determine any specific advantage of ANS based training. Due to their various pitfalls the other studies also could not convincingly demonstrate that ANS training had any specific benefits. The lack of clear results is in sharp contrast with how ANS studies are uncritically cited in the literature. We conclude that citation patterns reflect strong bias towards the ANS theory. Similar bias is reflected in study designs which avoid testing plausible and likely null hypotheses challenging the number sense theory and rather focus on unlikely or even in principle already rejected hypotheses. Hence, it is a plausible danger that the ANS training literature may develop into a highly cited 'null field' where null hypotheses are poorly formed and are posed in a way which biases them for rejection. In order to avoid this we suggested more optimal design options than used in the past and highlighted current errors. We owe delivering clear and unbiased information to children and their parents.

Tables

Wilson et al. (2006b); p2	*'the ability to represent and manipulate numerical quantities non-verbally'.*
Wilson et al. (2009); p224	*'the ability to quickly understand, approximate and manipulate numerical quantities'*
Räsänen et al. (2009); p452	*'Sense of approximate magnitudes'*
Obersteiner et al. (2013); p125.	*'…system represents larger numerosities approximately'*
Sella et al. (2016)	*This paper seems to use 'number sense' in the sense used by Jordan et al. (2012)*
Hyde et al. (2014); p92.	*ANS: 'primitive cognitive system for making quantitative judgements and decisions: the … ANS'*
Dewind and Brannon (2012); p1	*'…approximate number sense that allows us to estimate quantity without the use of symbols and language.'*
Park and Brannon, (2013); p1	*'… an Approximate number system (ANS) that allows them [humans] to represent quantities as imprecise, noisy mental magnitudes without verbal counting or numerical symbols'*
Park and Brannon, (2014); p188	*'…an intuitive understanding of number. Without counting or the use of symbols, we are able to estimate, compare, and mentally manipulate large numerical quantities.'*
Wang et al. (2016); p83	*'an intuitive, non-symbolic, approximate sense of number that is available prior to the onset of schooling… The ANS represents numbers in a noisy imprecise fashion…'*

Table 1. ANS definitions from the papers discussed. The studies are cited in the order of discussing them in this paper.

Citation	Age and Group N	Test type (df)	Power (D=0.3, 0.5, 0.8)		
Wilson et al. 2006	*7 to 9 year-olds* N=9 N=8 (1 excluded)	 matched t(8) matched t(7)	 0.13 0.11	 0.26 0.23	 0.56 0.50
Wilson et al. 2009	*4 to 6-year-olds* 53 = 27+26	matched t(26) matched t(25)	 0.31 0.30	 0.69 0.67	 0.97 0.97
Räsänen et al. 2009	*6.5 year-old children* 59 = 2×15+29	Indep. t(15+29-2); Ratio=29/15	0.15	0.34	0.69
Obersteiner et al. 2013	*6.9 year-old children* Children 147=35+39+39+34	Approximate vs. exact training group: Indep. t(35+39-2); Ratio = 39/35	0.25	0.56	0.92
Sella et al. 2016	*5.1 year-old children* 45=23+22 BUT widely varying numbers in actual analyses!	Max: Indep. t(2×20-2) Min: Indep. t(2×9-2)	0.15 0.09	0.34 0.17	0.69 0.36
DeWind and Brannon (2012)	*Adults* 20	Correlation (r=0.148; 0.243; 0.371)	0.10	0.19	0.39
Park and Brannon (2013)	*Adults* Exp. 1: 52=2×26 Exp. 2: 46=16+14+16	Indep. t(2×26-2) Indep. t(2×16-2)	0.19 0.13	0.42 0.28	0.81 0.59
Park and Brannon (2014)	*Adults* Exp. 1: 71=3×18+17	Indep. t(2×18-2)	0.14	0.31	0.65
Hyde et al. 2014	*Grade 1 children* Exp. 1: 96 = 4×24 Exp. 1: 48 = 2×24	Indep. t(2×24-2)	0.17	0.40	0.77
Wang et al. 2016	*5-year-old children* 40 = 4×10	Indep. t-test(2×10-2)	0.10	0.19	0.40

Table 2. Samples sizes and power of studies. The table presents citations to the studies, participant numbers, the test types for which power was computed with degrees of freedom (df) and the computed power values in the last three columns. For studies comparing condition and/or group means power is computed for matched and/or independent-sample t-tests (α=0.05; two-tailed). This is because pairwise comparisons between conditions and/or groups were of interest for all studies. Where group sizes in different analyses varied greatly due to exclusions minimum (Min.) and maximum (Max.) power values are computed. Otherwise the best possible power or the most relevant (Oberteiner et al. 2013) scenarios were computed. In the study with correlations r to D transformation was computed as $r = d / \mathrm{sqrt}(d^2 + a) \mid a=4$ (Borenstein et al. 2009). Power for t-tests was computed in Matlab using the sampsizepwr function, taking the into account the actual group sizes. Power for correlations was computed in GPower 3.1.9.2. (Faul, 2007). Indep. t. = Independent sample t-test.

A	B	C	D	E	F	G	H	I
Study	Math Improved	ANS Causal *Subset of B*	ANS Acuity Improved	Vague Supportive Comments	Description or Similar Paradigm	Critical comments or 'no effect'	Specific Critical Comments *Subset of G*	Any citation
Wilson 2006b	29 58%	13 26%	2 4%	3 6%	14 28%	2 4%	2 4%	50
Wilson 2009	20 71%	7 25%	1 4%	0 0%	3 11%	4 14%	2 7%	28
Räsänen et al. (2009)	26 49%	3 6%	3 6%	3 5%	21 4%	0 0%	0 0%	53
Obersteiner et al. (2013)	8 57%	4 29%	2 15%	1 7%	2 14%	1 7%	0 0%	14
Sella et al. (2016)	1 0%	0 0%	0 0%	0 100%	0 0%	0 0%	0 0%	1
Hyde et al. (2014)	24 83%	25 87%	2 7%	1 4%	1 3%	1 3%	1 3%	29
Dewind and Brannon (2012)	10 56%	9 50%	5 28%	0 0%	2 11%	1 5%	1 5%	18
Park and Brannon (2013)	36 80%	35 78%	3 7%	3 7%	1 2%	2 4%	1 2%	45
Park and Brannon (2014)	11 73%	10 67%	1 7%	0 0%	1 7%	2 13%	2 13%	15
Wang et al. (2016)	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%	0 0%	0
TOTAL % = x/253	165 65%	106 42%	19 8%	11 4%	45 18%	13 5%	9 4%	253 citations (85 papers)

Table 3. Summary of citations to the studies discussed. The top numbers in each row show the number of citations, the percentages below show the percentage of citations relative to the absolute number in Column I. There were 85 articles citing any of the studies discussed. Column I ('Any citation') states how many of these 85 articles cited a particular study for any reason. Columns A-H state how many of the 85 citing articles cited a particular study to support a particular claim. Columns C and D are subsets of column B. That is, the numbers in columns B, D-G add up to the numbers in column I (e.g. in row one: 25+6+0+2+17=50). The bottom row expresses citations in terms of the total number of citations (287). Content of columns A-H: **(A)** Citation of study. **(B)** Claim: There was improvement (transfer effect) in symbolic math ability. **(C)** Claim: Symbolic math improved and it was implied or stated that the improvement was causally related to ANS training. This is a subset of citations given in column B. **(D)** Claim: There was improvement in approximation ability or ANS acuity. **(E)** Vague positive statements about the cited study. **(F)** The cited study was simply described and/or mentioned for some reason. **(G)** Claim: Specific or non-specific critical comments about the cited study or claiming that there were no effects in the cited study. **(D)** Highly specific critical comments were made about the cited study. This is a subset of citations given in column G. **(I)** Total number of citations (see above). This is the sum of the citations in columns B and D-G as noted above. The literature used to locate these studies was conducted in May 2016.

Figure captions

Figure 1. Illustration of decision curves and accuracy outcomes for various w values. The figure is from Szűcs et al. (2013); author copyright.

Figure 2. Design options. (A) Improvement on an outcome measure depends on the overlap between representations and processes (RoPs) affected by training and those required by the outcome measure. Squares denote RoPs. The shaded squares mark the RoPs thought to be trained directly. The arrows point to other representation somehow also affected by the training. The filled circles mark all RoPs affected by the training. The thick dashed borders denote RoPs required by the outcome measure. **(B)** A simple design taking stimulus and task specificity into account. The test phase can test outcome measures related to all possible task/stimulus combinations or only select ones, e.g. only symbolic comparison and addition denoted by the shaded area in B2. For simplicity the figure does not represent pre, mid and post-test and other details explained in the text.

Finding number sense intervention studies

In May of 2016 an electronic literature search was conducted utilizing Google Scholar, Elsevier, PubMed, Scopus, and Web of Science search engines. Search criteria were that the papers should describe interventions studies which aimed to train the ANS with the intention of transferring training benefits to symbolic mathematics. Exact search terms used can be viewed in **Table A1**. From the initial hits, the titles were quickly scanned for appropriateness leaving 6,030 articles. Note that Google Scholar produced a large number of hits; consequently, the titles of the first 20 pages were looked at carefully as they were much more likely to be relavant while the subsequent webpages were scanned very quickly. The 6,030 articles were chosen as the titles seemed to have something to do with improvement in math ability or performance. The titles and abstracts of these were read to evaluate fit with the selection criterion from which 10 articles were specifically selected as ANS intervention studies. Additionally, the Introduction and Discussion sessions and the literature lists of all 10 articles were checked to see whether they cite other similar articles of interest. No other articles of interest were identified besides the initial 10 studies.

Search Terms
mathematics number sense intervention
mathematics number sense intervention review
arithmetic number sense intervention
number race
math ANS intervention
mathematics ANS intervention
math approximate number system intervention
mathematics approximate number system intervention
math magnitude representation intervention
mathematics magnitude representation intervention

math number sense training
mathematics number sense training
arithmetic number sense training
geometry number sense training
math ANS training
mathematics ANS training
arithmetic ANS training
geometry ANS training
math approximate number system training
mathematics approximate number system training
arithmetic approximate number system training
geometry approximate number system training
math magnitude representation training
mathematics magnitude representation training
arithmetic magnitude representation training
geometry magnitude representation training

Table A1. The search terms used in the literature search.

Appendix 2: Methods of the citation analysis

A search was conducted during May 2016 with the Elsevier and Web of Science search engines to find articles which cited the 10 ANS intervention studies. Eighty-six total citing articles were found. These articles were examined to determine what they concluded about the ANS intervention they were citing. First, the direct citations which discussed specific ANS intervention studies were found by searching within the document for the first author's last name of the intervention study in question. Second, the titles and abstract of the papers were read to see whether they had a critical stance to the papers discuss here. Third, the text of the papers was also checked for relevant critical comments. Based upon the text, we scored each citation in the citing papers along the criteria laid out in **Table 3**. The direct citations as well as information about what each concluded is available in the

Supplementary Material.

The citation data is available as an Microsoft Excel File published as **supplementary material** XX.

Legend for the supplementary Excel file: The file lists each paper which cites the 10 ANS intervention studies discussed here and codes them as follows: 2 = symbolic math competency/skills improved or shown to be causally based on ANS training; 1 = ANS acuity only improved; 0 = no effects shown or confounds in study; -1 = uses a similar paradigm, describes the paradigm, or tells what the study aims to do; -2 = vague supportive comments.

Appendix 3: Methods of effect size computation

Effect sizes were computed as defined by Hedges (1981):

$$G = \frac{m_1 - m_1}{SD}$$

Where m_1 stands for the mean performance score of study group 1, m_2 stands for the mean performance score of study group 2 and SD stands for the pooled standard deviation computed as:

$$SD = \sqrt{\frac{(n_1 - 1)sd_1^2 + (n_2 - 1)sd_2^2}{n_1 + n_2 - 2}}$$

Where sd_1 and sd_2 stands for the standard deviations measured in the groups and n_1 and n_2 denote the sample sizes in groups.

References

Ashcraft, M. H. (1982). The development of mental arithmetic: A chronometric approach. *Developmental Review, 2*, 213-36.

Boggan, M., Harper, S., & Whitmire, A. (2010). Using manipulatives to teach elementary mathematics. *Journal of Instructional Pedagogies, 3,* 1-6.

Borenstein M, Hedges LV, Higgins JPT, Rothstein HR (2009), *Introduction to Meta-analysis. Chapter 7.* John-Wiley and Sons. Ltd.

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365-76. doi:10.1038/nrn3475

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.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

de Castro, M. V., Bissaco, M. A. S., Panccioni, B. M., Rodrigues, S. C. M., & Domingues, A. M. (2014). Effect of a virtual environment on the development of mathematical skills in children with dyscalculia. *PLoS One, 9*(7), e103354.

Dehaene, S. (1997). *The number sense.* New York: Oxford University Press.

DeWind, N. K. & Brannon, E. M. (2012). Malleability of the approximate number system: Effects of feedback and training. *Frontiers in human neuroscience, 6*(68), 1-10. doi:10.3389/fnhum.2012.00068

Dyson, N. I., Jordan, N. C., & Glutting, J. (2013). A number sense intervention for low-income kindergartners at risk for mathematics difficulties. *Journal of Learning Disabilities, 46*(2), 166-81.

Ebersbach, M., Luwel, K., Frick, A., Onghena, P., and Verschaffel, L. (2008). The relationship between the shape of the mental number line and familiarity with numbers in 5- to 9-year old children: evidence for a segmented linear model. *J. Exp. Child Psychol.* 99, 1–17

Educational Testing Service (2016). Graduate Record Exam.

Educational Testing Service (2016). SAT.

Evans, S. H. & Anastasio, E. J. (1968). Misuse of analysis of covariance when treatment effect and covariate are confounded. *Psychological Bulletin, 69*(4), 225-34.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191

Fennema, E. H. (1972). Models and mathematics. *The Arithmetic Teacher, 19*(8), 635-40.

Fritz, C.,O., Morris, P.,E., Richler, J.,J, Effect size estimates: Current use, calculations and interpretation. *Journal of Experimental Psychology: General.* 141, 2-18 (2012).

Froebel, F. (1899). *Pedagogics of the Kindergarten: Ideas Concerning the Play and Playthings of the Child.*

Fuson, K. C. & Briars, D. J. (1990). Using a base-ten blocks learning/teaching approach for first- and second-grade place-value and multidigit addition and subtraction. *Journal for Research in Mathematics Education, 21*(3), 180-206.

Green CS, Strobach T, Schubert T (2014). On methodological standards in training and transfer experiments. Psychological Research. 78, 756-772

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6*(2), 107-28.

Hassinger-Das, B., Jordan, N. C., Glutting, J., Irwin, C., & Dyson, N. (2014). Domain-general mediators of the relation between kindergarten number sense and first-grade mathematics achievement. *Journal of Experimental Child Psychology, 118*, 78-92.

Hiebert, J. (1984). Why do some children have trouble learning measurement concepts? *The Artithmetic Teacher, 31*(7), 19-24.

Hoekstra, R. Morey, R.D., Rouder, J.N., Wagenmakers, E.J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin and Review. 21*, 1157-1164.

Hyde, D. C., Khanum, S., & Spelke, E. S. (2014). Brief non-symbolic approximate number practice enhances subsequent exact symbolic arithmetic in children. *Cognition, 131*(1), 92-107. doi: 10.1016/j.cognition.2013.12.007

Jordan, N. C., Glutting, J., Dyson, N., Hassinger-Das, B., & Irwin, C. (2012). Building kindergartners' number sense: A randomized controlled study. *Journal of Educational Psychology, 104*(3), 647-60.

Jordan, N. C., Glutting, J., & Ramineni, C. (2010). The importance of number sense to mathematics achievement in first and third grades. *Learning and Individual Differences, 20,* 82-8.

Jordan, N. C., Kaplan, D., Oláh, L. N., & Locuniak, M. N. (2006). Number sense growth in Kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development, 77*(1), 153-75.

Jordan, N. C., Locuniak, M. N., & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice, 22*(1), 37-47.

Kuhn, J. T. & Holling, H. (2014). Number sense or working memory? The effect of two computer-based trainings on mathematical skills in elementary school. *Advances in Cognitive Psychology, 10*(2), 59-67. doi:10.5709/acp-0157-2

Kuncel, N. R., Hezlett, S. A., & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. *Psychological Bulletin, 127*(1), 162-81.

Lindskog, M., Winman, A., & Juslin, P. (2013). Are there rapid feedback effects on approximate number system acuity? *Frontiers in Human Neuroscience, 7,* 1-8.

Lindskog, M., Winman, A. (2016). No evidence of learning in non-symbolic numerical tasks – A comment on Park and Brannon (2014). *Cognition*. 150, 243-251.

Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 68*(5), 304-5.

Lord, F. M. (1969). Statistical adjustments when comparing preexisting groups. *Psychological Bulletin*, 72(5), 336-37.

Marzola, E. S. (1987). Using manipulatives in math instruction. *Journal of Reading, Writing, and Learning Disabilities, 3*(1), 3-20.

Mazzocco, M. M. M., Feigenson, L., & Halberda, J. (2011). Impaired acuity of the approximate number system underlies mathematical learning disability (dyscalculia). *Child Development, 82*(4), 1224-37. doi: 10.1111/j.1467-8624.2011.01608.x

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 103-15.

Miller, G. A. & Chapman, G. A. (2001). Misunderstanding analysis of covariance. *Journal of Abnormal Psychology, 110*(1), 40-8.

Mönkkönen, A., Richardson, U., Räsänen, P., Herrera Montes, A., Kujala, J., Brem, S., et al. (in preparation). Graphogame-Math: Using a computer game for training number skills in preschool aged children.

Montessori, M. Translated by Everett, A. (1882). *The Montessori Method.* New York: Frederick A. Stokes Company.

Moreau D, Kirk IJ, Waldie KE (2016). Seven pervasive statistical flaws in cognitive training interventions. Frontiers in Human Neuroscience. 10:153

Moyer, R. S. & Landauer, T. K. (1967). Time required for judgements of numerical inequality. *Nature: Letters to Nature, 215*, 1519-20. doi:10.1038/2151519a0

Nath, S. & Szűcs, D. (2014). Construction play and cognitive skills associated with the development of mathematical abilities in 7-year-old children. *Learning and Instruction, 32*, 73-80.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan D., Kraut A., Lupia, A., Mabry, P., Madon, T. A., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E. L., Simonsohn, U., Soderberg, C., Spellman, B. A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E. J., Wilson, R., & Yarkoni, T. (2015b), Promoting an open research culture. *Science, 348*(6242), 1422-5.

Obersteiner, A., Reiss, K., & Ufer, S. (2013). How training on exact or approximate mental representations of number can enhance first-grade students' basic number processing and arithmetic skills. *Learning and Instruction, 23*, 125-35. doi:10.1016/j.learninstruc.2012.08.004

Park, J. & Brannon, E. M. (2013). Training the approximate number system improves math proficiency. *Psychological Science*. doi:10.1177/0956797613482944

Park, J. & Brannon, E. M. (2014). Improving arithmetic performance with number sense training: An investigation of underlying mechanism. *Cognition, 133*(1), 188-200. doi:10.1016/j.cognition.2014.06.011

Parsons, H. M. (1974). What happened at Hawthorne? *Science, 183*(4128), 922-32.

Porter, A. C. & Raudenbush, S. W. (1987) Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology, 34*(4), 383-92.

Räsänen, P., Salminen, J., Wilson, A. J., Aunio, P., & Dehaene, S. (2009). Computer-assisted intervention for children with low numeracy skills. *Cognitive Development, 24*(4), 450-72. doi:10.1016/j.cogdev.2009.09.003

Schmidt, F.L. (1992). What do data really mean? Research findings, meta-analysis and cumulative knowledge in psychology. *American Psychologist, 47*, 1173-81.

Sedlmeyer, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of the studies? *Psychological Bulletin. 105*, 309-16.

Sella, F., Tressoldi, P., Lucangeli, D., & Zorzi, M. (2016). Training numerical skills with the adaptive videogame "The Number Race": A randomized controlled trial on preschoolers. *Trends in Neuroscience and Education, 5*(1), 20-9. doi: 10.1016/j.tine.2016.02.002

Simmons, J., Nelson, L., & Simonsohn, U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science*, 22, 1359-66.

Sowell, E. J. (1989). Effects of manipulative materials in mathematics instruction. *Journal for Research in Mathematics Education, 20*(5), 498-505.

Suydam, M., & Higins, J. (1977). Activity-based learning in elementary school maththematics: Recomendations from research. Columbus, OH:ERIClearinghouse for Science, Mathematics, and Environmental Education. (ERICDocument ReproductionServiceNo.ED14840).

Szűcs, D., Devine, A., Soltesz, F., Nobes, A., & Gabriel, F. (2014). Cognitive components of a mathematical processing network in 9-year-old children. *Developmental Science, 17*(4), 506-24.

Szűcs, D., Nobes, A., Devine, A., Gabriel, F. C., & 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,* 1-12. doi: 10.3389/fpsyg.2013.00444

Van Dijck, J. P. & Fias, W. (2011). A working memory account for spatial-numerical associations. *Cognition, 119*(1), 114-9.

Wang, J. J., Odic, D., Halberda, J., & Feigenson, L. (2016). Changing the precision of preschoolers' approximate number system representations changes their symbolic math performance. *Journal of Experimental Child Psychology, 147*, 82-99. doi:10.1016/j.jecp.2016.03.002

Walsh, V. (2003). A theory of magnitude: Common cortical metrics of time, space, and quantity. *TRENDS in Cognitive Sciences, 7*(11), 483-8. doi:10.1016/j.tics.2003.09.002

White, S. L. J., Szűcs, D., & Soltész, F. (2011). Symbolic Number: Spatial representations in children aged 6 to 8 years. *Frontiers in Psychology, 2*(392), 1-11. doi: 10.3389/fpsyg.2011.003922

Wilson, A. J., Dehaene, S., Dubois, O., & Fayol, M. (2009). Effects of an adaptive game intervention on accessing number sense in low-socioeconomic-status kindergarten children. *Mind, Brain, and Education, 3*(4), 224-34. doi: 10.1111/j.1751-228X.2009.01075.x

Wilson, A. J., Dehaene, S., Pinel, P., Revkin, S. K., & Cohen, D. (2006a). Principles underlying the design of "The Number Race", an adaptive computer game for remediation of dyscalculia. *Behavioral and Brain Functions, 2*, 19-10.1186/1744-9081-2-19.

Wilson, A. J., Revkin, S. K., Cohen, D., Cohen, L., & Dehaene, S. (2006b). An open trial assessment of "The Number Race", an adaptive computer game for remediation of dyscalculia. *Behavioral and Brain Functions, 2*(1), 1-16. doi:10.1186/1744-9081-2-20

Zwick, R. & Sklar, J. C. (2005). Predicting college grades and degree completion using high school grades and SAT scores: The role of student ethnicity and first language. *American Educational Research Journal, 42*(3), 439-64.