Effects of Response and Presentation Format on Measures of Approximate Number System Acuity

Marcus Lindskog (marcus.lindskog@psyk.uu.se), Anders Winman (anders.winman@psyk.uu.se), Peter Juslin (peter.juslin@psyk.uu.se)
Department of Psychology, Uppsala University, P. O Box 1225,
Uppsala, 751 42 Sweden

Abstract

Human adults, infants, and non-human animals are believed to be equipped with an Approximate Number System (ANS) supporting non-symbolic representations of numerical magnitudes. Recent research has questioned both the validity and reliability of tasks intended to measure acuity in the ANS. Issues with validity and reliability might be due to differences in methodology. In the present study, we compare four tasks designed to measure ANS acuity, using a within-subjects design. The tasks are compared with respect to response and presentation format effects previously studied in the psychophysics literature, but largely ignored in the ANS literature. We find a presentation format effect and show that when non-symbolic numerical stimuli are presented sequentially the magnitude of the second stimulus is overestimated. Further, the results indicate that people's sensitivity to differentiate between non-symbolic numerosities is dependent on response format. The implications of the results to measures of ANS acuity are discussed.

Keywords: Approximate number system, response format, presentation format, validity, reliability

Introduction

Imagine walking in the countryside. As you approach a large field you spot two flocks of sheep, one with only white sheep and one with only black sheep, and amuse yourself by making a snapshot judgment of whether there are more white than black sheep. Later in your walk, you encounter another two flocks of sheep. This time the two flocks emerge from a tunnel, one flock after the other, separated by some short time interval. Once again, you test your judgment skills by deciding which of the two flocks that is the more numerous.

Human adults, infants and non-human animals have a common ability to represent numerical magnitudes, such as the number of sheep, without using symbols (Feigenson, Dehaene, & Spelke, 2004). The core cognitive system supporting this ability, the Approximate Number System (ANS), represents magnitudes in an approximate fashion with representations becoming increasingly imprecise as numerosity increases (Dehaene, 2009; but see, Brannon, Wusthoff, Gallistel, & Gibbon, 2001).

The accuracy with which the ANS can represent numerical magnitudes, often referred to as the acuity of the ANS, is conceptualized as the smallest change in numerosity that can be reliably detected and is often quantified by a Weber fraction (w). Acuity in the ANS progresses (i.e., w decreases) developmentally from childhood to adolescence (Halberda & Feigenson, 2008; Halberda, Ly, Wilmer, Naiman, & Germine, 2012) but even among adults there is considerable individual variability (e.g., Halberda & Feigenson, 2008; Halberda et al., 2012; Tokita & Ishiguchi, 2010).

Studies using brain imaging have identified a neurological basis for the ANS in the intraparietal sulcus (IPS) on the lateral surface of the parietal lobe (Castelli, Glaser, & Butterworth, 2006). Within the IPS, specialized neurons (numerons) sensitive to numerosity have been identified in macaque monkeys (Nieder, Freedman, & Miller, 2002). The IPS, however, is not only activated by non-numerical stimuli. In humans, it is also activated when they observe numbers in different modalities and when they perform simple arithmetic tasks (Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004). That the IPS is activated for a wide variety of numerical stimuli suggests a relationship between ANS acuity and achievement in formal mathematics. Such a relationship has been documented with children, even when controlling for a large number of cognitive abilities, (Halberda, Mazzocco, & Feigenson, 2008; Inglis, Attridge, Batchelor, & Gilmore, 2011) but results from studies on adults are mixed (Gebuis & van der Smagt, 2011; Inglis et al., 2011; Price, Palmer, Battista, & Ansari, 2012).

At least part of the mixed results might be attributed to differences in methodology. Recently the reliability and validity of some of the most commonly used measures of ANS acuity have been challenged (Gebuis & Van der Smagt, 2011; Gilmore, Attridge, & Inglis, 2011; Lindskog, Winman, Juslin, & Poom, 2013; Price et al., 2012) and studies indicate that while some formats show reasonable reliability and validity others are neither reliable nor valid (Lindskog et al., 2013). The differences in reliability and validity between different tasks that measure ANS acuity highlight the question of whether task features influences performance. Put differently, will you be better at deciding which of the two flocks of sheep that is the more numerous when you see them coming out of the tunnel, one flock at a time, or when you see both flocks at the same time on the field? The question of what factors that influence the reliability of ANS acuity measures is important also because reliability sets an upper limit on correlations between ANS acuity and other cognitive abilities. The present study addresses this question by comparing four ANS acuity tasks, in a within-subjects design, that use different presentation and response formats.

Response and Presentation Formats

While ANS acuity tasks use response and presentation formats that have been studied within the psychophysics literature (e.g. Macmillan & Creelman, 2005) little or no attention has been given to how the choice of format influences the measures of ANS acuity per se. The typical ANS acuity tasks present participants with two arrays of non-symbolic stimuli. Most often the stimuli are dots (e.g., Halberda et al., 2008) but other types of stimuli, for example arrays of squares, have also been used (e.g., Maloney, Risko, Preston, Ansari, & Fugelsang, 2010). After being presented with the two arrays participants' ability to differentiate between the numerosities of the two arrays is tested using one of two response formats. With a *comparison* format, similar to the two-alternative forced choice (2AFC) procedure used in psychophysics (Macmillan & Creelman, 2005), participants indicate which of the two arrays that is the more numerous. For example, Halberda et al. (2008) presented participants with two arrays of dots, one array of yellow dots and one of blue dots, and asked participants to indicate whether blue or yellow was the more numerous color. In contrast, with a *discrimination* format, similar to a same-different procedure used in classification tasks in psychophysics (Macmillan & Creelman, 2005), participants are to respond whether the two arrays have the *same* amount of dots or if the amounts of dots in the two arrays *differ*. The distinction between the two formats is relevant because even though psychophysicists have long known that the discrimination format is notoriously more difficult than the comparison format (Macmillan & Creelman, 2005) this difference has not been acknowledged and investigated within the ANS-literature. Therefore, the present study compares the two response formats directly in a within-subjects design.

In addition to the discrimination-comparison distinction, tasks that measure ANS acuity can be classified with respect to how the stimulus is presented temporally. With a *simultaneous* presentation format, both arrays of stimuli are presented at the same time. For example, in the study by Halberda et al. (2008) the arrays of blue and yellow dots were spatially intermixed and presented on a monitor at the same time. In contrast to the simultaneous presentation format, several studies have employed a *sequential* presentation format (e.g., Gilmore et al., 2011) where the two arrays are presented one at a time, separated temporally by a short interstimulus interval (ISI). Two reasons make the presentation format distinction important. First, previous research has indicated that while tasks using a simultaneous presentation format exhibit a reasonably good validity, tasks with a sequential presentation format do not (Lindskog et al., 2013). Second, the introduction of an ISI in 2AFC tasks has been shown to introduce a bias, the *time-order-error* (TOE), where the second stimulus is commonly judged larger more often than the first (Hellström, 1985; Macmillan & Creelman, 2005). Consequently, a sequence presenting the two arrays as; *Less numerous → More numerous*, would be correctly reported more often than the opposite sequence; *More numerous → Less numerous*. Whether a TOE exists or not in ANS acuity tasks is an empirical and potentially important question. In the present study, we compare the two presentation formats and investigate if a TOE is present in ANS acuity tasks when using the sequential presentation format together with the comparison response format.

The Present Study

Experiment 1 was designed to investigate the effects of presentation format and response format on performance in non-symbolic number differentiation tasks (i.e., ANS acuity tasks). To foreshadow the results; the experiment documented a TOE with a sequential presentation format and a comparison response format. Experiment 2 was designed to investigate the origin of the TOE by the use of direct estimates of non-symbolic numerosities. Experiment 1 also documented an effect of response format with better performance in the comparison than the discrimination format. We designed Experiment 3 to investigate if this effect was due to features of the ANS or due to a difference in sensitivity related to task features.

Experiment 1

In Experiment 1, participants performed four tasks designed to measure ANS acuity. The tasks were modeled from those used in previous research on ANS acuity. The experiment was designed to compare response formats and presentation formats in general and more specifically to investigate if two classical phenomena documented in the psychophysics literature, the time-order-error and the comparison/discrimination difference, were present in tasks measuring ANS acuity.

Method

Participants. Participants (10 Male, 20 Female) were undergraduate students from Uppsala University with a mean age of 26.1 years (*SD* = 6.6 years). They received a movie ticket or course credits for their participation.

Materials and procedure. Participants carried out a set of four tasks, described in detail below and illustrated in Figure 1. The order of tasks was counterbalanced using a Latin square. In none of the tasks did participants receive feedback on their performance.

Parallel comparison. The parallel comparison task was based on Halberda et al. (2008). On each of the 200 trials, participants saw spatially intermixed blue and yellow dots on a monitor. Exposure time (200ms) was too short for the dots to be serially counted. We used five ratios between the numerosity of the two arrays of dots (1:2, 3:4, 5:6, 7:8, 9:10) with the total number of dots varying between 11 and 30. One fifth of the trials consisted of each ratio. Half of the

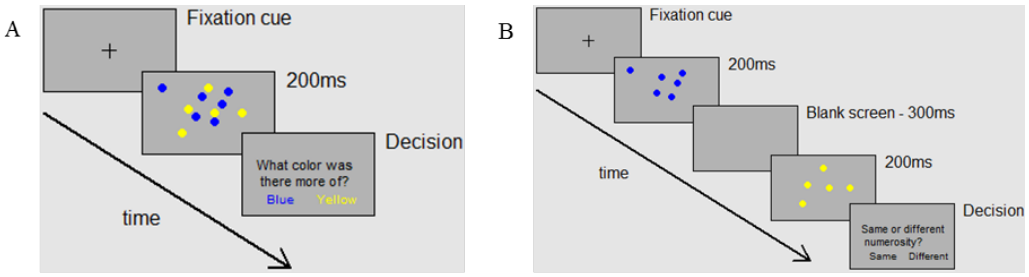

Figure 1: Illustration of the parallel (Panel A) and sequential (Panel B) presentation formats together with the comparison (Panel A) and discrimination (Panel B) response format.

trials had blue and half had yellow as the more numerous set. The dots varied randomly in size. To counteract the use of perceptual cues dot arrays were matched for total area on half of the trials and for average dot-size on the other half of the trials. The participants judged which set was more numerous by pressing a color-coded keyboard button.

Sequential comparison. The sequential comparison task used the same stimuli as the parallel comparison task. Here, however, the stimuli were presented sequentially and separated by a 300 ms interstimulus interval. The order of color, and whether the first or second array was the more numerous, was counterbalanced over trials.

Parallel discrimination. The parallel discrimination task presented the stimuli in the same way as the parallel comparison task. Stimuli for half of the trials were created as in the comparison tasks with the same ratios between the numerosity of the two sets of dots and the same total number of dots. For the second half of the trials, both sets of dots (i.e. the blue and yellow set) had the same number of dots. Using the same numerosities as when the two sets differed in the number of dots resulted in the total number of dots varying between 10 and 32. In addition, while the comparison tasks required participants to determine whether the blue or the yellow set of dots was the more numerate, the parallel discrimination task asked participants to determine if the two sets of dots had the *same* or *different* amount of dots.

Sequential discrimination. The sequential discrimination task used the same presentation format as the sequential comparison task and the same response format and stimuli as the parallel discrimination task.

Results and Discussion

Because the discrimination tasks do not easily allow for the modeling of an individual weber fraction, and because previous research (Lindskog et al., 2013) indicates that proportion correct is just as reliable and valid as *w*, we used proportion correct as a measure of performance in all of the four tasks.

We compared performance in the four tasks by entering proportion correct as dependent variable into a 2x2 repeated measures ANOVA with presentation format (parallel/sequential) and response format (comparison/discrimination) as within-subjects independent variables. This analysis showed a significant main effect of presentation format ($F(1,29) = 55.6$, $p < .001$) with better performance with the sequential format ($M = .73$, $SEM = .009$) than with the parallel format ($M = .67$, $SEM = .008$). There was also a significant main effect of response format ($F(1,29) = 546.1$, $p < .001$) with higher proportion correct with the comparison ($M = .81$, $SEM = .01$) than with the discrimination ($M = .59$, $SEM = .008$) format. The two-way presentation format by response format interaction did not reach significance ($F < 1$).There were no effects of the order of the ANS tasks ($F < 1$).

In the two tasks using a sequential presentation format, the dot arrays are presented in one of two orders, either the larger or the smaller array came first. To investigate if this ordering influenced performance we entered proportion correct as dependent variable into a 2x2 repeated measures ANOVA with response format (comparison/discrimination) and array-size order (larger-smaller/smaller-larger) as within-subjects independent variables. The significant interaction ($F(1,29) = 19.2$, $p < .001$), illustrated in Figure 2, show that while the array-size order does not influence performance with the discrimination format there is a significant and substantial difference between the two orders with the comparison format.

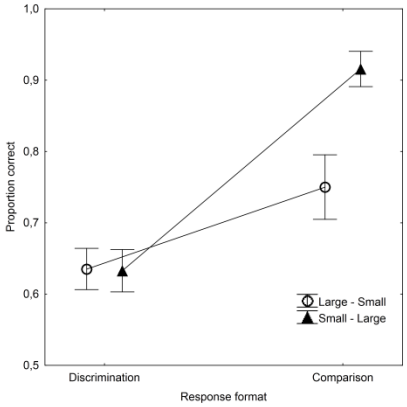

Figure 2: Proportion correct as a function of response format and array order. Vertical bars denote 95 % - confidence intervals.

In the comparison format the order which presents the smaller array first leads to significantly higher proportion correct (M =.92 , SEM = .012) than the order which presents the larger array first (M =.75 , SEM = .022). The analysis thus suggested that there might be a TOE present.

A TOE may occur either because the first stimulus is underestimated, because the second stimulus is overestimated or because it is psychologically easier to detect an increase in numerosity rather than a decrease. We designed Experiment 2 to distinguish between these three possibilities and to investigate the origin of the TOE when using non-symbolic numerosities as stimuli.

The response format effect might emerge for at least two different and independent reasons. First, it might be a feature of the ANS that it is adapted to detect the direction of a difference. For example, the ANS might have developed to determine that bush A contains *more* berries than bush B, rather than to just determine that there is a *difference* in the amount of berries on the two bushes. Second, it might be that participants' sensitivity is higher with the comparison format than with the discrimination format as suggested by signal detection theory (Macmillan & Creelman, 2005). We designed Experiment 3 to distinguish between these two possibilities.

Experiment 2

Experiment 2 investigated the origin of the TOE observed in Experiment 1. Participants made direct estimates of the number of displayed dots in a task closely matching the sequential tasks from Experiment 1.

Method

Participants. Twenty undergraduate students took part in the study, 12 females and 8 males. Average age was 24.8 (SD = 5.49). Participants received a cinema voucher or course credits for their participation.

Stimulus and procedure. Stimuli were three numbers of dots (8, 11, and 14) that were presented in temporal sequence in stimulus pairs (e.g. 8 – 11, 14 – 8 etc.) in a randomized order in a fully crossed design (two presentation positions (first/second) by three numerosities (8/11/14)). The dots were either blue or yellow. The sequence of colors was always the same for each participant, but randomized between subjects. Each stimulus pair was presented 9 times. Intermixed with these stimulus pairs each numerosity also occurred in isolation as a control. Together this made up 96 trials per participant. The numerosities were presented for 200ms, with a blank interstimulus interval of 300ms. Half of the trials were controlled for average dot-size, half for cumulative area. The task consisted of directly estimating the number of dots. This was done by entering a single number (for control stimuli) or two numbers with the keyboard. The input box was color coded, and always occurred in a left/right fashion corresponding to first/second position. Participants were told that if they altogether had missed a presentation of stimuli, they could indicate this by entering an error code.

Results and Discussion

Stimuli for which participants indicated that they had missed the presentation, as well as outlier responses ($|z| > 3$) were excluded from the analysis. These data made up 2.3% of the responses. There were no effects of color sequence order or stimulus type (size/area controlled).

Figure 3 shows judgments for control stimuli that appeared in isolated presentations. As can be seen in the figure, ratings were quite sensitive to the number of dots ($F(2, 38) = 38.0$, $p < .001$, one-way repeated measures ANOVA), but with a slight overestimation (the actual number is depicted in the dotted line in the figure).

Figure 4 shows the data of the two presentation positions and different numerosities. As can be seen in the figure, ratings are higher in the second presentation position.

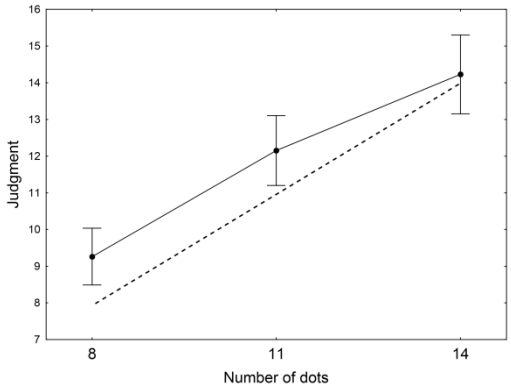

Figure 3: Mean judgments of the three numerosities when presented separately (dotted line shows actual numerosity). Vertical bars denote 95 % - confidence intervals.

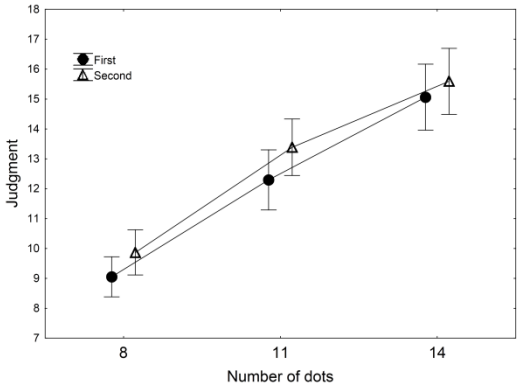

Figure 4: Mean judgments of the three numerosities for each presentation position (first /second). Vertical bars denote 95 % - confidence intervals.

A two-way ANOVA with numerosity (8/11/14) and stimulus presentation position (first/second) as independent within-subjects variables and judged numerosity as dependent variable shows that this presentation position effect is statistically significant ($F(1,19) = 34.5$, $p < .001$). The interaction was not significant ($F(2, 38) = 1.8$, $p = .17$)

A one-way ANOVA with condition (control/1st presentation/2nd presentation) as independent variable shows a significant effect on absolute error ($F(2, 38) = 11.4$, $p < .001$). Error was lowest in the control condition ($M = 1.93$, $SEM = .26$), higher in the first presentation position ($M = 2.16$, $SEM = .27$) and highest in the second presentation position ($M = 2.55$, $SEM = .34$). A Scheffé's post hoc test revealed that the error in the second presentation position was statistically significant from the two other conditions, which did not differ significantly from each other. The means of the absolute difference between participants' estimates and the three numerosities 8, 11, and 14 (i.e. the absolute error) were 1.42, 2.16, and 2.28 respectively. This increase in absolute error was statistically significant ($F(2,38) = 6.8$, $p = .00292$).

The results of Experiment 2 show that when two numerosities occur in a brief temporal sequence, separated by a short interval, the second numerosity is rated as more numerous than the first, and with a larger error. There is no clear indication of interference in the reversed temporal direction, presenting a second numerosity does apparently not have a deteriorating effect on the judgment of the first numerosity. The results support the one of the proposed explanations for the TOE found in Experiment 1; Participants' better performance with the smaller → larger presentation order than with the larger → smaller order is due to the inflation in experienced numerosity of the second stimulus. This effect leads to participants correctly identifying this order, but will hinder performance on the larger → smaller sequence.

Experiment 3

In Experiment 3, we added an extra response alternative to the comparison format. In addition to answering whether the blue or yellow array was the more numerous, participants could also respond that they had the same numerosity. If a feature of the ANS is that it is adapted to detect direction (i.e. to detect that A > B rather than just that A ≠ B) we expected better performance with this new response format than with a discrimination format as a result of increased performance on trials with different amount of dots. However, if the effect from Experiment 1 could be attributed to a change in sensitivity, the opposite was expected because adding the extra response alternative would make the task harder.

Method

Participants. The participants from Experiment 2 participated in Experiment 3.

Materials and procedure. Half of the participants carried out the two comparison tasks described in Experiment 1. The other half carried out the same tasks but with an alteration to the response format. The alteration combines the response formats of the comparison and discrimination tasks in Experiment 1. In the original discrimination tasks, participants could respond *same* or *different* while the comparison format had *blue* and *yellow* as response alternatives. In the modified task, participants had three response options: *blue*, *yellow*, and *same*. All other features of the task were identical to the comparison tasks of Experiment 1. The order of tasks was counterbalanced.

Results and Discussion

We compared performance in the four tasks by entering proportion correct as dependent variable into a 2x2 split-plot ANOVA with presentation format (parallel/sequential) as within-subjects independent variable and response format (same-different/yellow-same-blue) as between-subjects independent variable. Both the main effect of presentation format ($F(1,18) = 36.3$, $p < .001$) and the main effect of response format ($F(1,18) = 20.4$, $p < .001$) were significant while the interaction was not ($F < 1$). Participants in the same-different condition performed better ($M = .61$, $SEM = .012$) than did those in the yellow-same-blue condition ($M = .51$, $SEM = .012$). Further, and replicating the results from Experiment 1, performance was better in the sequential ($M = .60$, $SEM = .013$) than in the parallel presentation format ($M = .52$, $SEM = .011$).

These results show that the response format difference from Experiment 1 was eliminated, and even reversed, when an extra response alternative was added to the comparison format. This indicates that the discrimination format is more difficult than the comparison format and that the difference seen in Experiment 1 could be accounted for by a difference in sensitivity. However, even though the results lend support for a sensitivity explanation it does not exclude the possibility that the ANS is adapted not only to detect differences but also to detect the direction of a difference. This should be a question for future research to examine in more detail.

General Discussion

Recently, a large body of research has investigated the ANS and its relationship to mathematical achievement. This research has used several different tasks to measure ANS acuity. The present study extends previous research by investigating response and presentation format effects on performance in ANS acuity tasks and shows that comparisons between tasks might not always be straightforward.

In Experiment 1, we found three effects with potentially important implications. First, the sequential presentation format yielded approximately 8% (.72 vs. .67) better performance than the parallel format. In ANS experiments

were *w*, which is modeled on proportion correct (e.g., Halberda et al., 2008), rather than proportion correct is used as a performance measure this corresponds to a 30-45% difference in *w* for a typical participant (*w* = [.15 - .20]). Thus, changing the presentation format can give rise to a substantial difference in estimated *w*.

Second, in the sequential comparison task the order of stimulus was found to affect performance, similar to a TOE. Experiment 2 showed that the effect was due to an overestimation of the second stimulus compared to the control stimulus while no such bias could be found for the first stimulus. While it remains for future research to determine why the second stimulus is overestimated, one possibility could be residual activation in the IPS from the first stimulus. The effect has implications for measurements of ANS acuity. First, it will be necessary for future research using a sequential presentation format to counterbalance the order of stimulus for each ratio. Second, counterbalancing the order of stimulus might not be sufficient if numerosities are used that give rise to asymmetric differences in proportion correct. That is if the gain in one presentation order is larger/smaller than the loss in the opposite order. It remains for future research to investigate such asymmetries.

Finally, performance with a comparison format was significantly better than with a discrimination format. We proposed two possible explanations for this effect, that the ANS is adapted to detect direction or a difference in sensitivity, and showed in Experiment 3 that the latter was supported. This suggests that research on ANS acuity might benefit from, in addition to *w* and proportion correct, introducing a measure of sensitivity as a performance measure. The pattern of results, however, does not exclude the possibility of the ANS being a system adapted to detect the direction of a difference. This possibility should be an intriguing question for future research.

Acknowledgments

This research was sponsored by the Swedish Research Council. We thank Håkan Nilsson, Philip Millroth, Leo Poom, and Mona Guath for valuable comments on earlier versions of the manuscript.

References

Brannon, E. M., Wusthoff, C. J., Gallistel, C. R., & Gibbon, J. (2001). Numerical subtraction in the pigeon: Evidence for a linear subjective number scale. *Psychological Science*, *12*, 238–243.

Castelli, F., Glaser, D. E., & Butterworth, B. (2006). Discrete and analogue quantity processing in the parietal lobe: A functional MRI study. *Proceedings of the National Academy of Sciences of the United States of America*, *103*, 4693–4698.

Dehaene, S. (2009). Origins of Mathematical Intuitions. *Annals of the New York Academy of Sciences*, *1156*, 232–259.

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

Gebuis, T., & Van der Smagt, M. J. (2011). False approximations of the Approximate Number System? *PLoS ONE*, *6*, e25405.

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

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

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

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

Hellström, Å. (1985). The time-order error and its relatives: Mirrors of cognitive processes in comparing. *Psychological Bulletin*, *97*, 35–61.

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

Lindskog, M., Winman, A., Juslin, P. & Poom, L. (2013). *Measuring approximate number system acuity reliably*. In preparation.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: a user's guide* (2nd ed.). Mahwah, N.J: Lawrence Erlbaum Associates.

Maloney, E. A., Risko, E. F., Preston, F., Ansari, D., & Fugelsang, J. (2010). Challenging the reliability and validity of cognitive measures: The case of the numerical distance effect. *Acta Psychologica*, *134*, 154–161.

Nieder, A., Freedman, D. J., & Miller, E. K. (2002). Representation of the quantity of visual items in the primate prefrontal cortex. *Science*, *297*, 1708–1711.

Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, *44*, 547–555.

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

Tokita, M., & Ishiguchi, A. (2010). How might the discrepancy in the effects of perceptual variables on numerosity judgment be reconciled? *Attention, Perception & Psychophysics*, *72*, 1839–1853.