Acta Psychologica 150 (2014) 120–128

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/ locate/actpsy

Concurrent validity of approximate number sense tasks in adults and children

Karolien Smets [a,*], Titia Gebuis [a], Emmy Defever [a,b], Bert Reynvoet [a,b]

[a] Experimental Psychology, KU Leuven, 3000 Leuven, Belgium
[b] Faculty of Psychology and Educational Sciences, KU Leuven @ Kulak, 8500 Kortrijk, Belgium

ARTICLE INFO

Article history:
Received 25 October 2013
Received in revised form 30 April 2014
Accepted 5 May 2014
Available online 27 May 2014

PsycINFO codes:
23402260

Keywords:
Number processing
Change detection
Comparison
Same-different
ANS acuity
Validity

ABSTRACT

Reasoning with non-symbolic numerosities is suggested to be rooted in the Approximate Number System (ANS) and evidence pointing to a relationship between the acuity of this system and mathematics is available. In order to use the acuity of this ANS as a screening instrument to detect future math problems, it is important to model ANS acuity over development. However, whether ANS acuity and its development have been described accurately can be questioned. Namely, different tasks were used to examine the developmental trajectory of ANS acuity and studies comparing performances on these different tasks are scarce. In the present study, we examined whether different tasks designed to measure the acuity of the ANS are comparable and lead to related ANS acuity measures (i.e., the concurrent validity of these tasks). We contrasted the change detection task, which is used in infants, with tasks that are more commonly used in older children and adults (i.e., comparison and same-different tasks). Together, our results suggest that ANS acuity measures obtained with different tasks are not related. This poses serious problems for the comparison of ANS acuity measures derived from different tasks and thus for the establishment of the developmental trajectory of ANS acuity.

Humans, but also non-human species, are equipped with an Approximate Number System to estimate and compare different sets of items (ANS; Brannon, 2006; Cordes, Gelman, Gallistel, & Whalen, 2001; Feigenson, Dehaene, & Spelke, 2004; Libertus & Brannon, 2010). This system is rooted in the intraparietal area of the brain (Dehaene, Piazza, Pinel, & Cohen, 2003; Nieder, Freedman, & Miller, 2002; Piazza, Izard, Pinel, Le Bihan, & Dehaene, 2004; Sawamura, Shima, & Tanji, 2002) and represents numerosities in an approximate manner on a mental number line from left to right (Dehaene, 1997). Because of these approximate numerical representations, numerosities which are close to each other on the mental number line will overlap. This representational overlap between numerosities that are closer to each other (e.g., four and five) makes it more difficult to discriminate between them than between numerosities that are further apart (e.g., three and nine).

The numerical representations in the ANS become noisier with increasing numerosity, since there is more representational overlap between larger numerosities, (e.g., Dehaene, 1992; Gallistel & Gelman, 1992). This is in correspondence with the adherence of the ANS to Weber's Law (Fechner, 1860), stating that a proportionally larger difference between two numerosities is required with increasing magnitude in order to maintain a constant level of discrimination performance. Hence, when people have to judge which of two numerosities is larger, their performance is determined by the relative and not the absolute difference between the numerosities. Overall, performance is more accurate and responses are faster when the relative difference between the numerosities increases (Barth, Kanwisher, & Spelke, 2003; Defever, Reynvoet, & Gebuis, 2013; Price, Palmer, Battista, & Ansari, 2012). The smallest relative difference or ratio between two numerosities that can be discriminated above chance can be held as an indicator for the acuity of the ANS and it has been demonstrated that this ANS acuity increases with age (e.g., Halberda & Feigenson, 2008; Van Oeffelen & Vos, 1982).

Several researchers demonstrated a significant positive relationship between ANS acuity and (future) math ability (e.g., Halberda, Mazzocco, & Feigenson, 2008; Piazza et al., 2010, but for alternative results, see De Smedt & Gilmore, 2011; Sasanguie, De Smedt, Defever, & Reynvoet, 2012; Sasanguie, Göbel, Moll, Smets, & Reynvoet, 2012; Soltész, Szucs, & Szucs, 2010). Children that were more proficient at comparing different sets of numerosities had better scores on mathematical achievement tests. Consequently, it is suggested that ANS acuity can be used as a screening instrument to identify at an early age children who are at risk for future math problems (Gersten, Jordan, & Flojo, 2005; Piazza et al., 2010). Children that deviate in performance from what is expected of a child at that age might be at risk for developing mathematical deficiencies or dyscalculia.

However, to model the developmental trajectory of ANS acuity accurately and to detect deviations from normal development, a valid and

* Corresponding author at: KU Leuven, Experimental Psychology, Tiensestraat 102, bus 3711, 3000, Leuven, Belgium.
E-mail address: Karolien.smets@ppw.kuleuven.be (K. Smets).

http://dx.doi.org/10.1016/j.actpsy.2014.05.001
0001-6918/

reliable measure is needed. To date, many researchers operate under the implicit assumption that different tasks measure the ANS in the same way and they therefore for instance include results from different tasks in a single graph describing ANS development (see Fig. 4 in Halberda & Feigenson, 2008; Fig. 3 in Piazza et al., 2010). However, recent research suggests otherwise. For instance, several studies showed that differences in the way that numerosities are presented (e.g., sequential versus parallel presentation) or different task instructions (e.g., "indicate the larger" in the comparison task or "detect the difference" in the same-different task) can affect estimates of ANS acuity (Inglis & Gilmore, 2013; Price et al., 2012; Sasanguie, Defever, Van den Bussche, & Reynvoet, 2011; Smets, Gebuis, & Reynvoet, 2013).

Differences in methodology between tasks designed to measure ANS acuity are even more pronounced when one compares tasks used in infant studies (i.e., preferential looking tasks; e.g., Libertus & Brannon, 2010) with techniques applied to older children and adults (i.e., comparison and same-different tasks; e.g., Defever, Sasanguie, Vandewaetere, & Reynvoet, 2012; Sasanguie et al., 2011). For instance, Gebuis and van der Smagt (2011) contrasted the comparison task, a task commonly used in older children and adults, with an explicit version of a habituation task typically used to test infants (i.e., a detection task; e.g., Xu & Arriaga, 2007; Xu & Spelke, 2000). Participants in the comparison task were presented with two numerosities and were instructed to indicate the larger of those two numerosities. The same participants in the detection task were presented with a continuous stream of 12 dots and occasionally a deviant numerosity was presented. Participants were instructed to detect the deviant numerosity by pressing a key. The results showed that participants performed worse on the detection task than on the comparison task, which suggests that a task that is more similar to infant tasks is more difficult than a task more commonly used to investigate adults' ANS acuity. Consequently, infant performance and their associated ANS acuity may have been underestimated in previous research.

The comparability of ANS acuity tasks used in different age groups (i.e., infants versus older children and adults) and thus the concurrent validity of different tasks is recently receiving more attention in the literature (e.g., Starr, Libertus, & Brannon, 2013). This is also the topic investigated in the current study. Specifically, we compared performance on a change detection task, which is used in infants (e.g., Libertus & Brannon, 2010), with performance on the comparison and the same-different task, which are frequently used in older children and adults (e.g., Defever et al., 2013). A direct comparison will provide more insight in the concurrent validity of these different tasks.

The change detection task we used was an adapted version of the task used in the study of Libertus and Brannon (2010). In this study, infants were presented with two streams of numerosities of which one remained constant in numerosity (e.g., 16-16-16…), while the other stream alternated between two different numerosities (e.g., 8-16-8-…). The ratio between the numerosities that alternated in the changing stream was manipulated. Libertus and Brannon (2010) found ratio-dependent looking-times: Infants looked longer at the changing stream if they were in fact able to discriminate between the numerosities that alternated in this changing stream, dependent on the ratio between these numerosities. In the present study, we transformed the change detection task of Libertus and Brannon (2010) into an explicit task. Participants were now instructed to indicate which of the two presented streams changed in numerosity. This was done to allow comparing performance on this task with performance on the explicit comparison and same-different tasks.

The present study consists of three different experiments. Since an overt response was not present in the original task (Libertus & Brannon, 2010), we tested both a direct and delayed response condition for the change detection task in Experiment 1. In the direct response condition, participants were free to answer at any given time within the trial: during the presentation of the streams of stimuli or at the end of the trial. Participants in the delayed response condition could only answer at the end of the trial. The presence of a ratio-dependent effect would suggest that this task is suitable to study ANS processing. In Experiment 2, we administered the change detection, the same-different and the comparison task to adults. This would give insight in the concurrent validity of the three tasks. In Experiment 3, we administered the change detection task to primary school children. The inclusion of primary school age children will show whether this task is suitable to test the ANS in children. Furthermore, the administration of this task will provide us with a measure of ANS acuity across different ages, obtained with a task that is more compatible with the task that is used in infants. Additionally, we compared performance of primary school children on the change detection task to performance of age-matched children on the comparison and same-different task. The latter results were derived from a previous study.

1. Experiment 1

1.1. Method

1.1.1. Participants

Thirty participants participated in the change detection task with a direct response (mean age = 20 years, 16 female) and fifteen participants took part in the change detection task with a delayed response (mean age = 24 years, 13 female). Participants either received course credits or were paid for their participation in the experiment. The experiment was approved by the Ethical Committee of the Faculty of Psychology and Educational Sciences of the University of Leuven. All participants gave written informed consent for their participation.

1.1.2. Stimuli and procedure

Participants in both the direct and delayed response condition were presented with two streams of non-symbolic numerosities, one on the left and one on the right side of the screen (see Fig. 1). On each trial, the same numerosity was presented in one stream (e.g., 9 dots, 9 dots, 9 dots…), while two different numerosities alternated in the other stream (e.g., 9 dots, 18 dots, 9 dots…). The stimuli were dot patterns ranging from 8 to 35 (see Table 1) and were created with an adapted version of the program developed by Gebuis and Reynvoet (2011). With this program, stimuli are created whose sensory properties are uninformative about numerosity across trials. Consequently, an increase in numerical distance is not associated with an increase in sensory properties across trials. To this end, in half of the trials, the different visual cues (dot diameter, convex hull, contour length, aggregate surface and density) are congruent with number and in the other half of the trials the visual cues are incongruent with number. We manipulated the ratio (larger numerosity/smaller numerosity) by which the numerosities within the changing stream differed, resulting in five different ratio conditions: 1.2, 1.25, 1.5, 2 and 2.5. The smallest ratio was the most difficult and the largest was the easiest.

Each trial started with a red fixation cross that was presented for 1000 ms, followed by a green fixation cross that remained on the screen for 500 ms. Next, the two streams of dot patterns were presented. Each dot pattern remained on the screen for 500 ms and a black screen was displayed for 500 ms in between the dot patterns (see also Libertus & Brannon, 2010). The stimuli in both streams changed seven times during one trial (see Fig. 1). Participants in the direct response condition could either respond during the presentation of the streams or at the end of the trial when a question mark was displayed. Participants in the delayed response condition could only respond at the end of the trial when the question mark was displayed. In both conditions, participants were instructed to indicate the stream that changed in numerosity by pressing a key at the corresponding side. For each ratio condition, there were 4 possible number pairs, which were repeated 16 times, resulting in 64 trials per ratio condition (5 ratios * 4 number pairs * 16 trials = 320 trials in total). Participants were administered

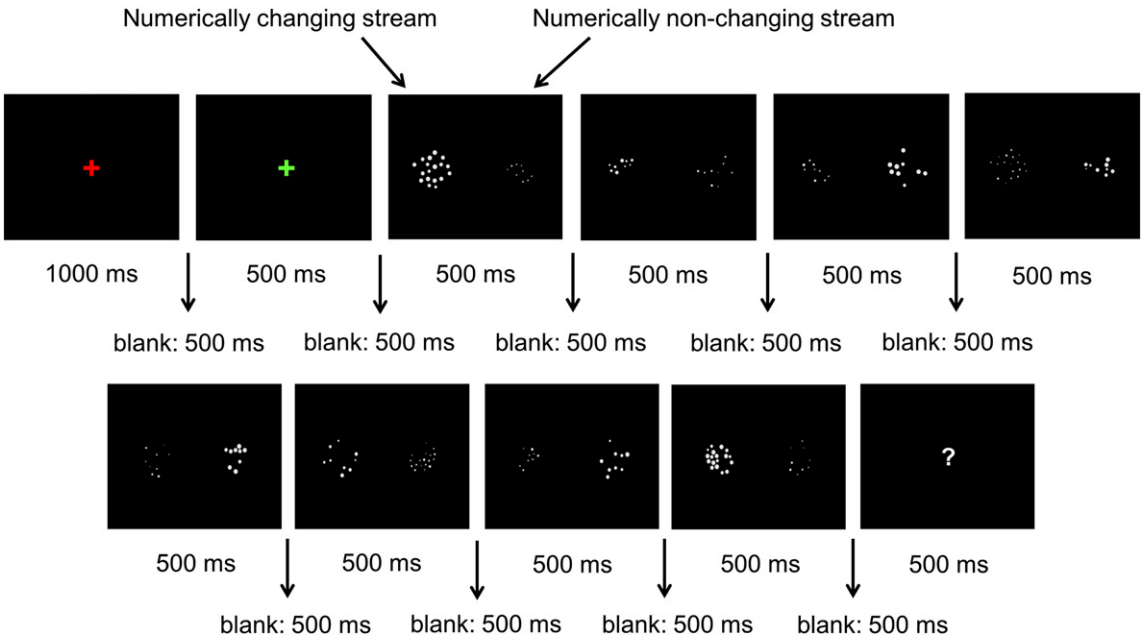

Fig. 1. Presentation of the stimuli in one trial of the change detection task.

10 practice trials before the experiment started and they could take a break in the middle of the experiment (after 160 trials).

1.2. Results

We calculated mean accuracies for each participant per ratio and included these in a repeated measures analysis with ratio (five levels: 1.2, 1.25, 1.5, 2 and 2.5) as a within-subjects factor and response instruction as a between-subjects factor (two levels: direct and delayed response). We included the factor response instruction in the analyses to investigate possible differences between the direct and delayed response conditions. We only focused on the accuracy scores and not on the reaction times for two reasons. First, the very large differences in reaction time between the direct and delayed response conditions that are due do the specific design of the respective conditions would make comparing them difficult and trivial. Second, previous research concerned with non-symbolic numerical cognition has frequently merely focused on accuracies instead of reaction times (e.g., Gebuis & van der Smagt, 2011; Gilmore, Attridge, De Smedt, & Inglis, 2014; Gilmore, Attridge, & Inglis, 2011; Halberda & Feigenson, 2008; Sasanguie, Defever, Maertens, & Reynvoet, 2013). When the assumption of sphericity was violated in the analyses, we corrected the p-values with the Greenhouse-Geisser correction (pGG).

The results for accuracy showed a significant main effect of response instruction, $F(1,43) = 17.85$, $p < .001$. Overall performance was better in the delayed response condition compared to the direct response condition (79% versus 71%). The main effect of ratio, $F(4,172) = 356.23$, $p < .001$, and the interaction between ratio and task instruction were significant, $F(4,172) = 3.05$, $pGG = .03$. Additional linear contrast analyses for both tasks separately indicated that participants in both response conditions became more accurate in detecting the changing stream when the ratio between the numerosities of the changing stream increased from 1.2 to 2.5 (direct response condition: $F(1,29) = 536.18$, $p < .001$; 55%, 57%, 71%, 85% and 88%; delayed response condition: $F(1,14) = 415.22$, $p < .001$; 59%, 62%, 81%, 95% and 97%). The interaction is indicative of a stronger ratio effect in the delayed response condition than in the direct response condition (see Fig. 2). Apparently, response instruction has an impact on the size of the ratio effect.

Table 1
Number Pairs.

Ratio	2.5	Ratio	2	Ratio	1.5	Ratio	1.25	Ratio	1.2
8	20	9	18	12	18	12	15	10	12
10	25	12	24	14	21	16	20	15	18
12	30	14	28	18	27	20	25	20	24
14	35	16	32	20	30	24	30	25	30

1.3. Discussion

The goal of Experiment 1 was to evaluate two different conditions of the explicit change detection task (Libertus & Brannon, 2010; Ross-sheehy, Oakes, & Luck, 2003) in adult participants: change detection with a direct response versus change detection with a delayed response. The results showed a significant ratio effect in accuracy for both response conditions, thus providing a measure of ANS acuity

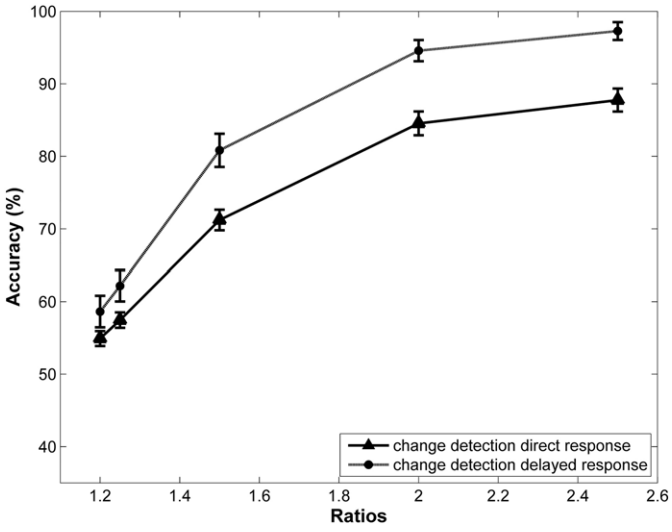

Fig. 2. Accuracy rates for the change detection task with a direct response and for the change detection task with a delayed response. A ratio effect in accuracy is present in both tasks, but performance is in general better and the ratio effect is stronger in the change detection task with a delayed response.

obtained with the explicit change detection task. More importantly however, there was an effect of response instruction on average performance. Participants in the delayed response condition performed better than participants in the direct response condition. Thus, when participants were required to take more time, they benefitted from this instruction and consequently performed better. Response instruction also played a role in the size of the ratio effect. Participants in the delayed response condition demonstrated a steeper ratio effect, probably due to the fact that they performed better, especially on the larger ratios.

In Experiment 2, we examined the validity of the change detection task, the comparison task and the same-different task by assessing whether performance of adults on these three tasks is comparable. This was achieved by exploring potential differences between the three tasks, but more importantly also by evaluating correlations among performances in the three tasks. The latter can provide conclusive evidence with regard to the comparability of the different tasks in Experiment 2 of the present study. More specifically, a significant correlation between performances on the tasks is expected if they are supported by some common mechanism. Because our focus in the present study is mainly on accuracy and because accuracy levels were highest in the change detection task with a delayed response, we opted to use the delayed response change detection task in the following experiments.

2. Experiment 2

2.1. Method

2.1.1. Participants

Twenty-two participants participated in Experiment 2 and performed three different number tasks (change detection with a delayed response, comparison, and same-different). They received course credits in exchange for their participation in the experiment. Three participants were excluded from further analyses because they made substantially more errors than the other participants (>2*SD*) in one of the three tasks. The final sample consisted of 19 participants (mean age = 19 years, 18 female). The participants gave written informed consent for their participation. The experiment was approved by the Ethical Committee of the Faculty of Psychology and Educational Sciences of the University of Leuven.

2.1.2. Stimuli and procedure

Participants were presented with three non-symbolic tasks: a change detection task with a delayed response, a comparison task and a same-different task. The change detection task was identical to the delayed response condition of Experiment 1. In the comparison task, participants were instructed to choose which of two simultaneously presented numerosities was numerically larger. In the same-different task, participants had to indicate whether two numerosities were numerically the same or different. The tasks were kept as similar as possible. We used the same number pairs and ratios, and chance level in all tasks was 50%. The stimuli and the ratios were the same as in Experiment 1. For both the change detection task and the comparison task, there were 4 possible number pairs that were repeated 8 times for each ratio condition, resulting in 32 trials per condition (5 ratios*32 trials = 160 in total).[1] For the same-different task, half of the trials consisted of two stimuli with the same numerosity and in the other half of the trials numerosity differed. The trials that differed in numerosity were the same stimulus pairs as those used for the change detection and comparison task. To keep the number of trials where the numerosities differed comparable between tasks, we added 40 "same" trials that were repeated 4 times to obtain 50% "same" trials and 50% "different" trials (160 "same" trials in total). Consequently, the analyses for all three tasks were performed on the same number of trials per ratio by excluding the "same" trials from the analyses. Thus, the change detection task and the comparison task consisted of 160 trials in total (32 per ratio), while the same-different task consisted of 160 "different" trials (32 per ratio) and 160 "same" trials (320 in total).

The procedure of the change detection task was identical to the procedure of the delayed response condition in Experiment 1 and thus participants had to postpone their response until the end of the trial. For both the comparison and the same-different task, each trial started with a red fixation cross that was presented for 500 ms. Next, the two numerosities were presented and remained on the screen for 1000 ms. They could either respond during the presentation of the numerosities or after they disappeared and a black screen was displayed. The order in which participants completed the three tasks was randomized. Participants were again administered 10 practice trials in each task and were told that they had the opportunity to take a break at certain moments during the task.

2.2. Results

For all tasks, mean accuracies were calculated for each participant per ratio. Similar as in previous studies (e.g., Sasanguie et al., 2011), the "same" trials of the same-different task were excluded from the analyses because these trials cannot be ascribed to a specific ratio condition. We therefore only included the "different" trials in the analyses. When the assumption of sphericity was violated in the analyses, we corrected the p-values with the Greenhouse-Geisser correction (pGG). Similar to Experiment 1, reaction times were not analysed. We performed a repeated measures analysis with accuracy as the dependent variable and both ratio (five levels: 1.2, 1.25, 1.5, 2 and 2.5) and task (three levels: change detection, comparison and same-different) as within-subjects factors.

The repeated measures analysis revealed a significant main effect of task, $F(2,36) = 20.56$, $pGG < .001$. Paired t-tests indicated significant differences in average accuracy between the comparison task (i.e., 79%) on the one hand and the change detection task (i.e., 68%), $t(18) = 5.15$, $p < .001$, and the same-different task (i.e., 60%), $t(18) = 7.02$, $p < .001$, on the other hand. The difference between the same-different and change detection task did not reach significance, although a trend towards a difference in accuracy between both tasks was present, $t(18) = 1.92$, $p = .07$. Separate pairwise t-tests for the different ratios between the change detection and same-different task were only significant for the largest ratio (i.e., 2.5; $t(18) = 2.73$, $p = .01$) and the smallest ratio (i.e., 1.2; $t(18) = 2.55$, $p = .02$). Participants performed better on the change detection task for these ratio conditions (Fig. 3). The main effect of ratio was also significant, $F(4,72) = 79.04$, $pGG < .001$, and was included in a significant interaction with task, $F(8,144) = 2.96$, $p = .04$. To ascertain that this interaction was not caused by the absence of a ratio effect in either of the tasks, a linear contrast analysis of performance on each ratio was conducted for all three tasks. These linear contrasts indicated that participants in all tasks became more accurate when the ratio between the numerosities increased from 1.2 to 2.5 (all F's > 27.43, all p's < .001; change detection: 59%, 53%, 56%, 81% and 89%; comparison: 63%, 68%, 79%, 91% and 95%; same-different: 47%, 53%, 59%, 70% and 73%). Therefore, the interaction seems to be caused by a difference in strength or slope of the ratio effect. The ratio effect is slightly more pronounced in the change detection and comparison task than in the same-different task (Fig. 3).

[1] The change detection task in Experiment 2 consisted of fewer trials (i.e., only half of the number of trials of the delayed change detection task in Experiment 1). To ascertain that fewer trials would not lead to different results, we calculated the average accuracy in the first and second block of the delayed change detection task of Experiment 1 (block 1 = 78%; delayed block 2 = 80%) and computed the correlation between accuracies in these two blocks. A significant correlation between average accuracies in the first and second block was present (r = .74, p = .001). Additionally, no significant difference between accuracies in the two blocks was present. We can therefore conclude that the change detection task in Experiment 2 is comparable to the change detection task in Experiment 1, even though the task in Experiment 2 consisted of half the number of trials compared to the task in Experiment 1.

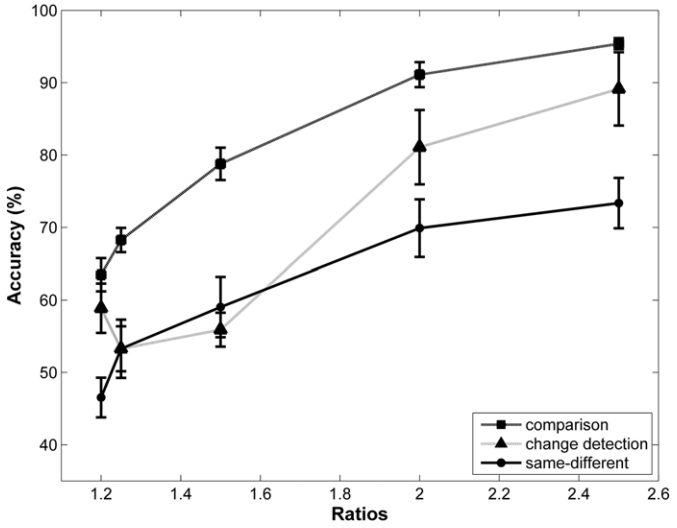

Fig. 3. Accuracy rates for the comparison task, the same-different task and the change detection task in adults. There was a significant ratio effect in all three tasks, although there were significant differences in average accuracy between the tasks which were associated with differences in the strength of the ratio effects.

Additionally, we calculated participants' Weber fractions in the three tasks by fitting a psychometric function on their accuracies (Wichmann & Hill, 2001). We fitted a cumulative Gaussian and corrected the model for potential biases (i.e., guess-rate and lapse-rate). Both guess-rate and lapse-rate were included as free parameters in the model. The Weber fraction was subsequently derived from the Weber ratio, while taking into account the participant's individually estimated guess-rate and lapse-rate. Six participants were excluded because of an extremely bad fit of the psychometric function on either one of the three tasks. The remaining Weber fractions were subjected to a repeated measures analysis with task (three levels: change detection, comparison and same-different) as a within-subjects variable. The analysis revealed a significant main effect of task, $F(2,30) = 8.49$, $pGG = .004$. Pairwise t-tests indicated significant differences between the Weber fractions of the comparison task ($w = 0.37$) on the one hand and the same-different ($w = 1$), $t(15) = 3.80$, $p = .002$, and change detection task ($w = 0.65$), $t(15) = 2.86$, $p = .01$, on the other hand. The difference in Weber fraction between the latter two tasks was only marginally significant, $t(15) = 1.93$, $p = .07$. These Weber fractions might seem high compared to other studies, but as Szücs, Nobes, Devine, Gabriel, and Gebuis (2013) already indicated, this may be due to constructing the stimuli with the script of Gebuis and Reynvoet (2011) which controls visual cues very stringently.

To further investigate the comparability of the different tasks, Pearson correlations between the average accuracies on all tasks were computed. The correlation between accuracy on the change detection task and accuracy on the same-different task, $r = -.05$, $p = .84$, and the correlation between accuracy on the comparison and accuracy on the same-different task, $r = .15$, $p = .53$, were not significant. However, the correlation between accuracy on the change detection task and accuracy on the comparison task did reach significance, $r = .52$, $p = .02$, suggesting that performance on both tasks is related. We verified whether the significant correlation between performance on the comparison and on the change detection task was not driven by outliers and therefore calculated Cook's distances. The cut-off values suggested by Bollen and Jackman (1990) were used. Participants with a value larger than 1 for Cook's distance were excluded from the analyses, as well as participants with a value larger than 4/n (n = number of participants) when exclusion of these participants changed the results of the correlational analysis. According to these cut-offs, two participants were excluded from the data and the correlation between accuracy on the change detection task and accuracy on the comparison task did not reach significance anymore, $r = .001$, $p = .99$. This indicates that these two participants exclusively caused the correlation between the change detection and comparison task (Fig. 4). Therefore, we can conclude that there is no significant correlation between accuracy on the change detection task and accuracy on the comparison task.

To verify whether the absence of significant correlations between average accuracies in the three tasks was not due to a lack of power in the data, we calculated the correlation between the first block of trials and the second block of trials in each task (i.e., the split-half reliability). These were all significant, all r's > .88 and all p's < .001, indicating sufficient power to detect a correlation when merely using half of the trials compared to the number of trials used to compute the correlations between the three tasks. Additionally, computing correlations between the Weber fractions of the three tasks also did not indicate a significant relationship between the different tasks (comparison and same-different: $r = -.28$, $p = .29$; comparison and change detection: $r = .41$, $p = .11$; change detection and same-different: $r = -.17$, $p = .53$).

2.3. Discussion

In our second experiment, we investigated whether performance on different non-symbolic number tasks is related. To address this issue, we administered three non-symbolic number tasks (the delayed change detection task, the comparison task and the same-different task) to the same adult participants.

The results of Experiment 2 showed the presence of a significant ratio effect in accuracy in all tasks, which is in accordance with previous studies (e.g., Defever et al., 2012; Halberda et al., 2008; Sasanguie et al., 2011). This ratio effect was of different strength across number tasks. The ratio effect of the change detection task and the comparison task was steeper than the ratio effect of the same-different task. Together, these results suggest that each non-symbolic number task measures the ANS, but that differences between the tasks exist (see also Price et al., 2012 for a similar conclusion).

Not only the strength of the ratio effect differed between the tasks, significant differences in average accuracy between the tasks were also present, and these results were supported by the Weber fraction analysis. Participants performed on average significantly better on the

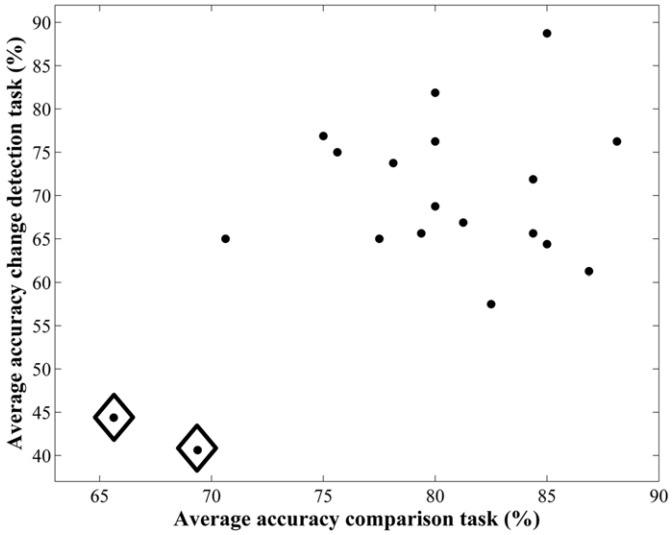

Fig. 4. Relationship between accuracy on the comparison task and accuracy on the change detection task. The diamond shapes indicate the outliers that solely caused the significant correlation. Without these outliers, the correlation between accuracy on the comparison task and accuracy on the change detection task did not reach significance anymore.

comparison task (i.e., higher accuracy and a smaller Weber fraction) than on the change detection and same-different task. The observation that a comparison task is easier than a detection task is in agreement with previous research (Gebuis & van der Smagt, 2011; Piazza et al., 2004). The difference in performance between the comparison task on the one hand and the same-different and change detection task on the other hand suggests that different mechanisms or processes lie at their basis (e.g., Smets et al., 2013; Verguts, Fias, & Stevens, 2005). The finding that the task that is most frequently used to investigate adults' numerical cognition is easier than a task that is more similar to infant tasks might have consequences for previous research that implicitly operated under the assumption that all these tasks measure ANS acuity in a similar way (e.g., Halberda & Feigenson, 2008; Piazza et al., 2010).

The difference in accuracy between the change detection task and the same-different task on some ratios and the marginally significant difference in Weber fraction between these two tasks indicated that participants performed slightly better on the change detection task than on the same-different task. This is related to the fact that the ratio effect was steeper in the change detection task than in the same-different task. The ratio effect in the same-different task is very small because performance on this task is quite low in general (Defever et al., 2013; Piazza et al., 2004). The task is too difficult for participants to perform better, especially on the larger ratios.

While significant differences in average accuracy between the tasks do not necessarily indicate that performances on these tasks are incomparable, the absence of a significant correlation does. If there is some common mechanism behind the tasks investigated in the present study, a highly significant correlation is expected. The correlation analysis can therefore provide conclusive evidence with regard to the comparability of the different tasks. However, our results indicated that performance on either of the tasks was not related to performance on the other tasks for both accuracy and Weber fractions, thus questioning the validity of the tasks. These absent correlations were not due to a lack of power and thus pertinently highlight that the tasks administered in the present study might not be as comparable as initially assumed. This is problematic considering that the implicit assumption behind some figures provided in some papers in which results from studies with different tasks are combined, is that all tasks are interchangeable tasks to investigate ANS acuity (e.g., Halberda & Feigenson, 2008; Piazza et al., 2010). As Inglis and Gilmore (2014) indicated, this requires the assumption that Weber fractions from different tasks with different methodologies are in fact entirely comparable. When assuming this however, at least a significant correlation between averages accuracies and Weber fractions on these tasks is expected. However, Starr et al. (2013) did find a significant relationship between 6-month-old infants' performance (i.e., a preference score by means of looking time) on the original implicit change detection task (Libertus & Brannon, 2010) and their performance on a comparison task at 3.5 years of age. Participants in their study however were very young children, and as Gilmore et al. (2014) indicated, ANS acuity tasks may be more related in children than in adults. Most ANS tasks entail a substantial amount of domain-general demands (e.g., working memory or executive function), which are especially demanding in very young children. Gilmore et al. (2014) considered it a possibility that ANS acuity tasks refer more to these domain-general task demands in children than in adults and that interference of these task demands can therefore be responsible for the significant correlation between different ANS acuity tasks in children, but not in adults.

In Experiment 3, we administered the change detection task to primary school children to assess whether the typical ratio effect and the classic developmental trend present in other number tasks (i.e., increasing accuracy with increasing age) was also evident when investigating numerical cognition with the change detection task. In the discussion, we compared performance on the change detection task of the present study with performance of similarly aged primary school children on a comparison and same-different task of another study.

3. Experiment 3

3.1. Method

3.1.1. Participants

Thirty-four children from the first grade (18 males, mean age = 6.71 years), 26 children from the second grade (16 males, mean age = 7.62 years), 25 children from the third grade (14 males, mean age = 8.13 years) as well as 25 children from the sixth grade (13 male, mean age = 11.43 years) participated in Experiment 3. Two children from the sixth grade were omitted from the analyses because they performed significantly worse than the other children from that grade (>2*SD*). The final sample thus consisted of 108 participants in total. The parents of the children were fully informed about the purpose and means of the study and gave written informed consent for their children's participation. The experiment was approved by the Ethical Committee of the Faculty of Psychology and Educational Sciences of the University of Leuven.

3.1.2. Stimuli and procedure

The children in the present study were presented with the same change detection task as in the previous experiment with adults, except for the specific number ratios. The ratios included in the study with adults would exceed children's numerical capacities. Therefore, we replaced the most difficult ratio (i.e., 1.2) of the adult study with a larger ratio of 1.75. We reasoned that including this ratio would a) make the task less difficult for children and hence keep them motivated, and b) make it possible to determine more precisely the limits of their numerical reasoning skills. The other ratios in the experiment with children were the same as in the adult experiment. Dot patterns ranged from 8 to 35 and were again created with the adapted version of the program from Gebuis and Reynvoet (2011). Four possible number pairs were repeated 8 times for each of the 5 ratio conditions, resulting in 32 trials per condition (5 ratios * 32 trials = 160 trials in total). To ascertain that all children understood the instructions appropriately, the experimenter showed and carefully explained 10 practice trials while indicating the relevant information as well as the correct answer. The procedure of the change detection task was identical to the procedure of the delayed response condition in Experiment 1. We included one break in the middle of the experiment.

3.2. Results

Mean accuracies in the change detection task were calculated for each participant per ratio. When the assumption of sphericity was violated in any of the following analyses, we corrected the p-values with the Greenhouse-Geisser correction (pGG). Similar as in Experiment 1 and Experiment 2, reaction times were not taken into account. First, we executed a repeated measures analysis with accuracy as the dependent variable, ratio (five levels: 1.25, 1.5, 1.75, 2 and 2.5) as a within-subjects factor and grade (four levels: first, second, third and sixth grade) as a between-subjects factor.

The repeated measures analysis showed a significant main effect of grade, $F(3,104) = 24.33$, $p < .001$. Independent t-tests however indicated that only the difference in performance between the oldest children (i.e., sixth graders with an average performance of 72%) and the rest of the children was significant (first grade: 55%, second grade: 57%, third grade: 59%; all t's > 4.50 and all p's < 0.001). The other t-tests showed that there were no significant differences between the remaining grades, all t's < 1.78 and all p's > 0.08. The main effect of ratio was also significant, $F(4,416) = 48.26$, $p < .001$. A linear contrast analysis verified that children became more proficient in indicating the changing stream when the ratio between the two numerosities that alternated in this changing stream increased from 1.25 to 2.5 (51%, 55%, 61%, 64% and 65%; $F(1,107) = 73.05$, $p < .001$). Grade and ratio were also involved in an interaction, $F(12,416) = 6.88$, $p < .001$. Separate linear

contrast analyses for each grade showed that there was no ratio effect in the first grade, $F(1,33) = 1.98$, $p = .17$. While there was no significant ratio effect present in the first grade however, the separate linear contrasts provided evidence for a linear increase in performance on the task with increasing ratio for the other grades, all F's > 9.93 and all p's < .004. To clarify whether it was the absence of a ratio effect in the first grade that caused the interaction, we performed the repeated measures analysis with ratio within-subjects and grade between-subjects again, but excluded the participants from the first grade. The interaction between ratio and grade remained significant, $F(8,284) = 4.23$, $pGG < .001$, indicating that the interaction between ratio and grade was not merely caused by the absence of a ratio effect in the first graders, but also by differences in the size of the ratio effect in the other grades (see Fig. 5).

3.3. Discussion

In the third experiment, we administered the change detection task to primary school children to examine whether this task can be used to obtain a measure of ANS acuity in these age groups. The results indicated that children from the second, third and sixth grade exhibited the expected ratio-dependent performance, but this ratio effect in accuracy was not present in the first graders. The figure of the respective ratio effects suggests that children from the first grade perform at chance level on nearly all ratios. This cannot be due to the specific ratios that were manipulated in the change detection task, considering that children of the same age are able to discriminate between numerosities that differ by these ratios in for instance the comparison task (Defever et al., 2013). Apparently, it is the specific instruction of the change detection task that makes the task too difficult for them.

In correspondence with other studies that used the comparison task or the same-different task (e.g., Defever et al., 2013; Halberda & Feigenson, 2008), we observed that children's accuracy increased with age when using the change detection task. However, only the difference in average performance between the sixth grade (i.e., mean accuracy of 79%) and the other grades (i.e., mean accuracies between 55 and 59%) reached significance, while the rest did not significantly differ. This finding suggests that only the sixth graders were actually able to perform the change detection task at a reasonable level, while the younger children struggled with it. This is especially the case for the first graders as evidenced above, but the task also seems to be fairly difficult for the second and third graders, as they perform around chance level on the most difficult ratios (1.5 and 1.25). Our findings further suggest that children from the first grade performed at the same level or even worse than the infants that participated in the implicit change detection task in the original study (Libertus & Brannon, 2010) considering that they performed around chance level on a ratio of 1.5 (where 9-month-old infants succeeded at discriminating between numerosities) or 2 (where 6-month-old infants succeeded at discriminating between numerosities). Children from the second and third grade performed only slightly better than infants, as illustrated by an above chance performance on ratio 1.5 and 2. Potential reasons for the worse performance of children in the present study, which are older than the infants in the study of Libertus and Brannon (2010), may lie in the fact that we used a different way of controlling the visual cues of the stimuli (i.e., the method as proposed by Gebuis & Reynvoet, 2011) which may make the task more difficult. Another potential explanation is that an implicit task may be a more sensitive measure of ANS acuity than an explicit task is and therefore leading to higher performances.

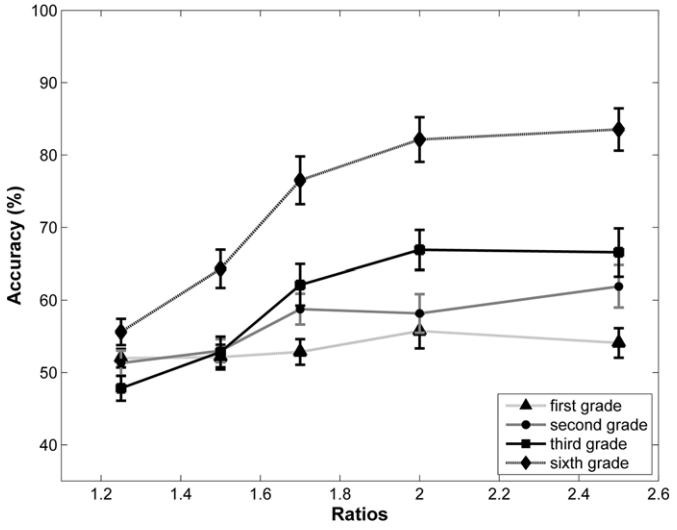

Fig. 5. Accuracy rates for the change detection task in different grades of primary school. The figure illustrates that there was no significant ratio effect in the first grade, while a ratio effect was present in the other grades. Children in the sixth grade performed significantly better than children in the other grades.

Recently, Defever et al. (2013) examined the performance of children of approximately the same age as in our study in a comparison and a same-different task. To compare our findings with the study of Defever et al. (2013), we re-calculated the average accuracy for the comparison task and the same-different task of their study and for the change detection task in the current Experiment 3, but only included those ratios that were identical in all tasks. A comparison of the children's performance on all three tasks revealed a slightly different pattern than in our experiment with adults (i.e., Experiment 2). For all grades, performance on the comparison task was the best (first grade: 71%, second grade: 85%, third grade: 84% and sixth grade: 94%). This was further confirmed by pairwise t-tests for each grade, all t's > 4.70 and all p's < .001. However, whereas the adult participants performed worst on the same-different task, this was not the case in primary school children. Their performance on the change detection task (first grade: 54%, second grade: 56%, third grade: 60% and sixth grade: 73%) and the same-different task (first grade: 56%, second grade: 60%, third grade: 64% and sixth grade: 68%) was more similar, as indicated by pairwise t-tests for each grade, all t's < 1.69 and all p's > .08. Clearly, the different tasks that are used to examine numerosity processing in both adults and children are not necessarily comparable, at least not in average performance.

4. General discussion

An accurate description of the developmental trajectory of ANS acuity is necessary to be able to detect children who are at risk for future mathematics problems at an early age (Gersten et al., 2005; Jordan et al., 2007; Piazza et al., 2010). Up till now, it is unclear whether the developmental trajectory of ANS acuity is accurately described, because different tasks were used both within and between age groups (e.g., Defever et al., 2013; Libertus & Brannon, 2010; Sasanguie, Göbel, et al., 2012; Xu & Spelke, 2000). It can be questioned whether concurrent validity is present for the different tasks that are used to obtain ANS acuity measures (Gebuis & van der Smagt, 2011; Gilmore et al., 2011; Price et al., 2012; Smets et al., 2013). Therefore, the main focus in the current study was to examine the validity of different tasks tapping into ANS acuity and to determine which task characteristics may be responsible for differences in ANS acuity. More specifically, we wanted to compare performance on tasks used in infants (i.e., the change detection task; Libertus & Brannon, 2010) with performance on tasks that are most commonly used in adults and older children (i.e., the comparison and same-different task).

In Experiment 1, two different conditions of the change detection task were evaluated (direct versus delayed response condition) to ascertain that a measure of ANS acuity can be obtained by applying this task to adults as an explicit task. A significant ratio effect for accuracy in the direct and delayed response conditions showed that indeed this task is suitable to measure ANS acuity. The results further indicated a

better performance in the delayed response condition than in the direct response condition. Based on these results, we decided to use this version of the change detection task in the following experiments.

In Experiment 2, performance of adult participants on the change detection task, the comparison task and the same-different task was contrasted. The results showed that both tasks might not be as comparable as initially assumed (e.g., Halberda & Feigenson, 2008; Piazza et al., 2010). There were differences in average accuracy and Weber fractions between the three tasks, which indicated that participants performed best in the comparison task, intermediate on the change detection and worst on the same-different task. Even more importantly, there was no significant correlation between the average accuracies or Weber fractions of the three tasks and this was not due to a lack of power in the data considering the significance of the correlations between performances in the two blocks of each experiment. Therefore, the results of Experiment 2 suggest that the implicit assumption in previous research (Halberda & Feigenson, 2008; Piazza et al., 2010) that all these tasks measure ANS acuity in the same way is probably unfounded.

Experiment 3 confirmed that this conclusion is partly applicable to primary school children of different grades. Overall, children performed worst on the same-different task and on our adapted version of the change detection task, while they performed best on the comparison task.

The difference in performance present in both adults and children and the observation that there was no significant correlation between the comparison task on the one hand and the change detection and same-different task on the other hand can be explained by the influence of decisional mechanisms. Using computational modelling, Verguts et al. (2005) proposed that performance on the symbolic comparison task is (partly) dependent on decisional mechanisms. More specifically, the connection weights between the relevant numerosity and the correct response node increase with increasing numerosity, which improves performance on the comparison task. However, as Van Opstal and Verguts (2011) indicated, decisional mechanisms only become evident in tasks where participants are instructed to indicate the larger or smaller numerosity (i.e., the comparison task). Hence, these decisional mechanisms do not play a role in the same-different task or the change detection task where the instructions refer to detecting a difference, rather than indicating the larger numerosity (Cohen Kadosh, Brodsky, Levin, & Henik, 2008; Smets et al., 2013). Considering that the decision principle of the three tasks remains the same in the non-symbolic task variants, the role that decisional mechanisms play in the different tasks may provide an explanation for our results. More specific, the fact that decisional mechanisms influence performance in the comparison task only may be the reason for the incomparability of the tasks and their unrelated performances.

Whereas both children and adults performed better on the comparison task than on the change detection or the same-different task, differences could also be observed between the latter tasks. The change detection task and the same-different task do seem to lead to different and unrelated performances in adults, but not in children where performance is more similar between both tasks. A first potential reason for the difference in performance and the non-significant correlation between the same-different and the change detection task in adults can be found in signal detection theories (Kingdom & Prins, 2009; Macmillan & Creelman, 2005). According to these theories, participants use a 'differencing strategy' in the same-different task to decide whether two stimuli are numerically different or not (i.e., subtracting the representations from the numerosities from each other and obtaining a perceived difference), because they are constrained by their approximate number skills to perform the same-different task (Macmillan, Kaplan, & Creelman, 1977). However, in order to actually make a judgment, each participant needs to determine a response criterion for this perceived difference. When the perceived difference exceeds this criterion, they will respond 'different', but when it does not, they will respond 'same'. Depending on whether participants set a criterion in which more (or less) evidence is required to respond 'different', a bias to responding 'same' (or 'different') can be induced. For example, if more evidence is necessary to reach a 'different' decision, a participant's performance will be lower compared to a participant who does not need that much evidence to reach this decision. The study of Defever et al. (2013) indicated that children have a bias towards responding 'different' in the same-different task, for instance because they may have difficulties ignoring the irrelevant visual cues that accompany numerosities. Their data additionally indicated that this bias decreases with age. Consequently, with age, participants require more evidence to reach a 'different' decision and become stricter in their responses. Thus, adults performing the same-different task need more evidence to reach a 'different' decision compared to children who revert to a 'different' decision more easily. The change detection task on the other hand is most probably criterion-free, considering that there is a difference present in each trial in either one of the streams (see also a same-different task with two pairs, which is also criterion-free; Kingdom & Prins, 2009). Therefore, there cannot be a bias present in the change detection task, which leads to a performance that is not affected by this bias especially present in children. Hence, although the same-different task and the change detection task both refer to detecting differences between numerosities, there are still important differences between them that can lead to differences in ANS acuity measures and unrelated performances, at least in adult participants.

Additionally, signal detection theory can offer a tentative explanation for the significant differences in performance between the comparison task and the same-different task and the absence of a correlation between them as well. Performance on the comparison task is also criterion-free, which leads to a fairly large difference in performance between the comparison task and the criterion-dependent same-different task (see also the Supplemental Data from Piazza et al., 2004). However, both the comparison and change detection task are criterion-free, but participants still seem to perform better on the comparison task. As illustrated above, this may have to do with specific characteristics of the comparison task (i.e., the role of decisional mechanisms; Verguts et al., 2005). Within the framework of signal detection theory however, performance on the comparison task is still differently influenced than performance on the change detection task even though both tasks are criterion-free. Participants will be able to use their approximate skills (Kingdom & Prins, 2009; Macmillan & Creelman, 2005) in the comparison task, while a differencing strategy is still the only option in the change detection task. This may make the change detection task more difficult, but still easier than the same-different task because participants are certain that there is in fact a difference in each trial. Thus, the fact that tasks are criterion-free or criterion-dependent may lead to different strategies that are used to perform the task, which in turn causes significantly different and unrelated performances.

A second potential reason which might clarify the unrelated and significantly different performances in the change detection task and the same-different task present in adults, but not in children, relates to the obvious differences in presentation duration or number of stimuli that are presented in both tasks. In the change detection task, participants were presented with two streams of each eight stimuli in every trial. Thus, they had several opportunities to detect the change. Participants in the same-different task needed to base their decision on whether there was a change in numerosity on only one pair of numerosities (i.e., only two stimuli were presented per trial). It therefore appears likely that this difference in stimulus presentation caused the rather small difference in performance. With the presentation of multiple stimuli per trial, there were more opportunities for adults to detect the difference in the change detection task than in the same-different task and this might have enhanced performance on the change detection task. Our results are therefore in accordance with but also extend the results of Price et al. (2012) who illustrated the importance of stimulus presentation in different variants of the comparison task. Additionally, our results also correspond to the results of Inglis and Gilmore

(2013), who suggested that stimulus presentation time (and thus in general, the time the participant observes the stimuli) plays an important role and are in correspondence with Experiment 1 of the present study. However, considering that children's working memory resources and executive functioning skills are less developed than in adults (Myatchin & Lagae, 2013), they might not benefit from the availability of more stimuli as is the case in the change detection task. On the contrary, they may have trouble with processing all the number information that is presented simultaneously in the change detection task. Therefore, this can be a second reason for the observation that the better performance on the change detection task than on the same-different task was not present in children.

In summary, the results from the present study suggest that a task that is more similar to a task used in infant studies (Libertus & Brannon, 2010) is not comparable to other tasks that are used in studies with older children and adults. As a side note however, it is important to note that in infants the mode of response is implicit, whereas in the present study an explicit response was required. It is possible that implicit and explicit measures of one and the same task still differ. Regardless, caution is necessary when evaluating the developmental trajectory of ANS acuity using different tasks, considering the lack of validity across these different tasks that are assumed to measure the same processes. The present study further suggests that the use of one and the same task in different age groups will provide a better and more valid tool to describe the entire developmental trajectory. It is therefore necessary to develop a standard method or task to measure ANS acuity and to avoid using different tasks to measure the same construct. Clear instructions with regard to the design of the task and a consensus on how to measure ANS acuity will provide more valid research and might enable us to clarify some opposite results from previous studies (e.g., on the presence or absence of a relationship between mathematics and ANS acuity). Another possibility is to explicitly evaluate the effects of different tasks and their characteristics on ANS acuity to provide more insight in the mechanisms that actually influence this ANS acuity.

Acknowledgements

The research was supported by the Research Fund KULeuven. Titia Gebuis was supported by the Marie Curie Intra-European Fellowship.

References

Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition*, *86*(3), 201–221.
Bollen, K. A., & Jackman, R. W. (1990). Regression diagnostics: An expository treatment of outliers and influential cases. *Modern methods of data analysis* (pp. 257–291).
Brannon, E. M. (2006). The representation of numerical magnitude. *Current Opinion in Neurobiology*, *16*(2), 222–229.
Cohen Kadosh, R., Brodsky, W., Levin, M., & Henik, A. (2008). Mental representation: What can pitch tell us about the distance effect? *Cortex*, *44*(4), 470–477.
Cordes, S., Gelman, R., Gallistel, C. R., & Whalen, J. (2001). Variability signatures distinguish verbal from nonverbal counting for both large and small numbers. *Psychonomic Bulletin & Review*, *8*(4), 698–707.
De Smedt, B., & Gilmore, C. K. (2011). Defective number module or impaired access? Numerical magnitude processing in first graders with mathematical difficulties. *Journal of Experimental Child Psychology*, *108*(2), 278–292.
Defever, E., Reynvoet, B., & Gebuis, T. (2013). Task- and age-dependent effects of visual stimulus properties on children's explicit numerosity judgments. *Journal of Experimental Child Psychology*, *116*(2), 216–233.
Defever, E., Sasanguie, D., Vandewaetere, M., & Reynvoet, B. (2012). What can the same-different task tell us about the development of magnitude representations? *Acta Psychologica*, *140*(1), 35–42.
Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, *44*(1), 1–42.
Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. Oxford University Press.
Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology*, *20*(3–6), 487–506.
Fechner, G. G. (1860). *Elemente der Psychophysik. Report in James (1890)*, *2*. (pp. 50). Leipzig: Breitkopf Und Hartel, 50.
Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, *8*(7), 307–314.
Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*(1), 43–74.
Gebuis, T., & Reynvoet, B. (2011). Generating nonsymbolic number stimuli. *Behavior Research Methods*, *43*(4), 981–986.
Gebuis, T., & van der Smagt, M. J. (2011). False approximations of the approximate number system? *PLoS ONE*, *6*(10), e25405.
Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities*, *38*(4), 293–304.
Gilmore, C., Attridge, N., De Smedt, B., & Inglis, M. (2014). Measuring the approximate number system in children: Exploring the relationships among different tasks. *Learning and Individual Differences*, *29*, 50–58.
Gilmore, C., Attridge, N., & Inglis, M. (2011). Measuring the approximate number system. *The Quarterly Journal of Experimental Psychology*, *64*(11), 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*(5), 1457.
Halberda, J., Mazzocco, M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, *455*(7213), 665–668.
Inglis, M., & Gilmore, C. (2013). Sampling from the mental number line: How are approximate number system representations formed? *Cognition*, *129*(1), 63–69.
Inglis, M., & Gilmore, C. (2014). Indexing the approximate number system. *Acta Psychologica*, *145*, 147–155.
Jordan, N. C., Kaplan, D., Locuniak, M. N., & Ramineni, C. (2007). Predicting first-grade math achievement from developmental number sense trajectories. *Learning Disabilities Research & Practice*, *22*(1), 36–46.
Kingdom, F. A. A., & Prins, N. (2009). *Psychophysics: A practical introduction*. Academic Press.
Libertus, M. E., & Brannon, E. M. (2010). Stable individual differences in number discrimination in infancy. *Developmental Science*, *13*(6), 900–906.
Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. New York, USA: Lawrence Erlbaum Associates.
Macmillan, Neil A., Kaplan, H. L., & Creelman, C. D. (1977). The psychophysics of categorical perception. *Psychological Review*, *84*(5), 452.
Myatchin, I., & Lagae, L. (2013). Developmental changes in visuo-spatial working memory in normally developing children: Event-related potentials study. *Brain and Development*, *35*(9), 853–864.
Nieder, A., Freedman, D. J., & Miller, E. K. (2002). Representation of the quantity of visual items in the primate prefrontal cortex. *Science*, *297*(5587), 1708–1711.
Piazza, M., Facoetti, A., Trussardi, A. N., Berteletti, I., Conte, S., Lucangeli, D., et al. (2010). Developmental trajectory of number acuity reveals a severe impairment in developmental dyscalculia. *Cognition*, *116*(1), 33–41.
Piazza, M., Izard, V., Pinel, P., Le Bihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human intraparietal sulcus. *Neuron*, *44*(3), 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*(1), 50–57.
Ross-sheehy, S., Oakes, L. M., & Luck, S. J. (2003). The development of visual short-term memory capacity in infants. *Child Development*, *74*(6), 1807–1822.
Sasanguie, D., De Smedt, B., Defever, E., & Reynvoet, B. (2012). Association between basic numerical abilities and mathematics achievement. *British Journal of Developmental Psychology*, *30*(2), 344–357.
Sasanguie, D., Defever, E., Van den Bussche, E., & Reynvoet, B. (2011). The reliability of and the relation between non-symbolic numerical distance effects in comparison, same-different judgments and priming. *Acta Psychologica*, *136*(1), 73–80.
Sasanguie, D., Göbel, S. M., Moll, K., Smets, K., & Reynvoet, B. (2012). Approximate number sense, symbolic number processing, or number-space mappings: What underlies mathematics achievement? *Journal of Experimental Child Psychology*, *114*(3), 418–431.
Sasanguie, D., Defever, E., Maertens, B., & Reynvoet, B. (2013). The approximate number system is not predictive for symbolic number processing in kindergartners. *The Quarterly Journal of Experimental Psychology*, *67*(2), 271–280.
Sawamura, H., Shima, K., & Tanji, J. (2002). Numerical representation for action in the parietal cortex of the monkey. *Nature*, *415*(6874), 918–922.
Smets, K., Gebuis, T., & Reynvoet, B. (2013). Comparing the neural distance effect derived from the non-symbolic comparison and the same-different task. *Frontiers in Human Neuroscience*, *7*, 28.
Soltész, F., Szucs, D., & Szucs, L. (2010). Relationships between magnitude representation, counting and memory in 4-to 7-year-old children: A developmental study. *Behavioral and Brain Functions*, *6*(1), 13.
Starr, A., Libertus, M. E., & Brannon, E. M. (2013). Number sense in infancy predicts mathematical abilities in childhood. *Proceedings of the National Academy of Sciences*, 201302751, http://dx.doi.org/10.1073/pnas.1302751110.
Szücs, 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*, http://dx.doi.org/10.3389/fpsyg.2013.00444.
Van Oeffelen, M. P., & Vos, P. G. (1982). A probabilistic model for the discrimination of visual number. *Perception & Psychophysics*, *32*(2), 163–170.
Van Opstal, F., & Verguts, T. (2011). The origins of the numerical distance effect: the same-different task. *Journal of Cognitive Psychology*, *23*(1), 112–120.
Verguts, T., Fias, W., & Stevens, M. (2005). A model of exact small-number representation. *Psychonomic Bulletin & Review*, *12*(1), 66–80.
Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293–1313.
Xu, F., & Arriaga, R. I. (2007). Number discrimination in 10-month-old infants. *British Journal of Developmental Psychology*, *25*(1), 103–108.
Xu, F., & Spelke, E. S. (2000). Large number discrimination in 6-month-old infants. *Cognition*, *74*(1), B1–B11.