Psicológica (2014), 35, 423-440.

A Functional Measurement Study on Averaging Numerosity

Michael D. Tira[*], Mariaelena Tagliabue, and Giulio Vidotto

University of Padua, Italy

In two experiments, participants judged the average numerosity between two sequentially presented dot patterns to perform an approximate arithmetic task. In Experiment 1, the response was given on a 0–20 numerical scale (categorical scaling), and in Experiment 2, the response was given by the production of a dot pattern of the desired numerosity (numerosity production). The experiments found that the responses were shaped according to an averaging integration model. This suggests the linearity in the response scale of both of the response methods in the approximate arithmetic task.

Numerosity, or "number sense", is a deeper understanding of the meaning of numbers that occurs without the ability to count and without any knowledge of numerical symbols. This ability has been studied for many years. Binet (1890) was the first to report about numerosity. He informally investigated the ability of children to compare the numerosity of two presented collections of simple objects. Binet (1890/1969, p. 87) concluded: "if [the child] judges one group more numerous than another, it is because it occupies more space on the paper". In 1929, De Marchi was the first to use a numerical evaluation of collectivities. According to De Marchi, the evaluation of collectivities refers to "the process by which a perceived aggregate is expressed by numerals in conditions that exclude any possibility of numbering its elements" (De Marchi, 1929/1986, p. 184). De Marchi acknowledged that variables influencing numerical evaluation—such as the duration of exposure, size of the surface, occupied by the single collectivities, density of the exposed elements (dots), or space and time disposition—could together influence the evaluation in an experiment. Years later, studies on these same variables that were used by De Marchi

[*] Address correspondence to Dr. Michael Tira, michael.d.tira@gmail.com

confirmed that the perceptual system is not able to abstract numerosity from other stimulus attributes (see Allik, 1989).

A variety of studies have demonstrated that non-human animals, including rats, lions, and various species of primates, have an approximate sense of numbers (for a review, see Dehaene, 1997). Human adults are able to estimate and manipulate approximate numerical magnitudes as well; this ability appears to be independent from language or other symbol systems, as it is present both in infants (Feigenson, Dehaene, & Spelke, 2004) and in non-human animals (Dehaene, 1997). The ability seems to be grounded in a general approximate system for magnitude representation, including senses of spatial extent and duration (Lourenco & Longo, 2011). Those senses have in common the ability to work across sensory modalities (e.g., vision and audition) and to share a “more vs. less” representational structure.

Since this paper focuses on arithmetic operations, it will be necessary distinguish between symbolic and non-symbolic aspects of elementary arithmetics. As defined by Dehaene (2009), “Symbolic arithmetic deals with how we understand and manipulate numerals and number words” (p. 233). Moreover, Dehaene (2009) states, “Nonsymbolic arithmetic is concerned with how we grasp and combine the approximate cardinality or “numerosity” of concrete sets of objects (such as visual dots, sounds, and actions)” (p. 233). The present work focuses on non-symbolic arithmetic.

The way that approximate numerical magnitudes are manipulated in order to judge (as opposed to calculate) the result of an arithmetic operation can be conceptualized as a multi-attribute judgment, with which the result is derived from the integration of the operands with a specific integration rule. This conceptualization allows for the application of the tools of Information Integration Theory (IIT)(Anderson, 1981; Anderson, 1982) to the study of mental arithmetic. Busemeyer (1991) summarizes some of the applications of IIT to the problem of intuitive estimations of algebraic operations on symbolic quantities (numbers) and continuous quantities (line lengths, tones, or weights). Moreover, in the field of IIT, many works use functional measurements to assess numerosity (Cuneo, 1982; Shanteau, Pringle & Andrews, 2007). Interestingly, no study has yet applied this approach to the investigation of the way in which the results of arithmetic operations with discrete quantities are computed or approximated. Thus, in the present work, two experiments were carried out to test the applicability of the IIT approach to approximate mental arithmetic of discrete quantities through the evaluation of the shape of the response function and of the goodness of fit of the model to behavioral data. We decided to use two different response methods to support the generality of the result.

IIT describes the psychological processes underlying multi-attributes decision-making and proposes a general method that is applicable to several contexts. IIT proposes a theoretical framework (cognitive algebra), which is accompanied by a methodology (functional measurement) that is relevant to the evaluation of its adaptation to the real contexts of the proposed models. IIT conceives the cognitive processes that lead to the integration of more information in a single concept (from physical stimuli *S* to a behavioral response *R*), as divided into three phases: evaluation, integration, and response. Each of these phases is governed by a specific function ($s=V(S)$, $r=I(s_1,s_2, \ldots, s_n)$, $R=M(r)$). This evaluation process leads to the assignment of an implicit value *s* to the individual constituent parts of the stimulus *S*. This is followed by an integration of these values that, in turn, leads to the formulation of an overall judgment. At this level, the different models that describe the ways the operation of integration is performed play a crucial role. The cognitive algebra framework provides three models of the integration process: the additive model (Anderson, 1962), the multiplicative model (Anderson & Shanteau, 1970; Anderson & Weiss, 1971), and the weighted average model (Anderson, 1965; Norman, 1976), also known as averaging. Thus, from this perspective, we can consider an algebraic operation as a process of the evaluation of a stimulus *S*, in which the operands are the constituent parts (S_1, S_2) of that stimulus. From this point of view, the process of evaluation includes the assignment of internal and subjective values, s_1 and s_2 to S_1 and S_2. This is followed by an integration of these internal values with an integration function. This leads to the formulation of an overall judgment, which represents the result of the algebraic operation. The functional measurement theory includes, besides each *s* value, a weight parameter. The weight represents the importance, assumed by the particular attribute in the overall judgment, and it is indicated by the parameter *w* in the models. Despite the fact that the theoretical formulation implies a distinction between scale values and weights, in both the additive and multiplicative models, the two parameters are not really distinguishable (Anderson, 1981). The effect of each attribute cannot be separated into a scale value and a weight. Conversely, the averaging model has the capability, under specific conditions, to distinguish between scale values and weights (Zalinski & Anderson, 1989).

The averaging model of IIT represents the subject's response to a multi-stimulus situation, as a weighted average. Each stimulus has two parameters: the weight *w*, which conveys the importance of the stimulus on the final judgment, and the scale value *s,* which represents its position on the dimension of response (Zalinski & Anderson, 1991). The averaging model represents the integrated response, *r,* as:

$$r = \frac{\sum w_t s_t}{\sum w_t}, \qquad t = 1, 2, ... \tag{1}$$

whereas, in a two stimuli situation, becomes:

$$r = \frac{w_1 s_1 + w_2 s_2}{w_2 + w_2} \tag{2}$$

The weight-value representations are common, but they are arbitrary in most formulations. Each weight in a standard regression model, for example, is confounded with the unit of the scale. The averaging model makes weight mathematically identifiable, and the empirical success of the model makes it psychologically meaningful (Zalinski & Anderson, 1991). The averaging model assigns weight and scale values to each stimulus. If all of the levels of one factor have the same weights $w_{Ai} = w_A$, then the model is said to be *equally weighte*d; if at least one of the levels differs, then the model is said to be *differently weighted.* Functional measurement makes use of the joint manipulation of at least two factors, according to a factorial design; the second block of each experiment was carried out for this purpose. From now on, we will refer to this as the factorial design block. Moreover, to differentiate the averaging model from the additive and multiplicative models, one or more factors at a time must be excluded from the factorial design; this is called a sub-design. The first experimental block of each experiment was meant explicitly for this purpose. From now on, the first block will be referred to as a sub-design block.

EXPERIMENT 1: CATEGORICAL SCALING

The aim of Experiment 1 was to study the integration rule, involved in approximate averaging operations of discrete quantities, and to evaluate the goodness of the fit of the averaging model to the data. First, we presented dot sets to participants and instructed them to indicate the numerosity of the sets on a 0–20 numerical scale. Later, we asked the participants to indicate on the same 0–20 scale the average numerosity between two sequentially presented dot sets. To test the integration rule that was involved in the task, we varied the number of presented dots systematically in a factorial design and in sub-designs. If the participants responded on a linearly distributed scale, and if they used an averaging integration rule to evaluate the averaging numerosity, then we would expect to find that the plot of the complete factorial design was a bundle of parallel lines, along with lines that represent sub-designs, intersecting the bundle.

METHOD

Participants. Fifteen undergraduate female students from the University of Padua participated in the experiment. The average age of participants was 21.5 years (*SD* = .5). A convenience sampling was used, and the participants received no payment.

Apparatus. Participants used a keyboard and a computer screen in a quiet room. The distance between the subject and screen was 70 cm. A Python program was developed in order to process the input from the keyboard and to control the presentation of stimuli. Importantly, the spatial pattern of the appearance of the dots was unpredictable. Precisely, with every .6 degree of clockwise rotation, one additional dot (2 mm in diameter, .16° of visual angle) was presented at a randomly chosen free position within an unmarked circular target area of 140 mm in diameter (11.42° of visual angle), centered on the screen. The minimum distance between the two dots was .25 mm (.02° of visual angle).

Materials. The random dot patterns were presented in white on a black background. A circular gray area with a radius of 140 mm was presented to the participants just before the dot pattern, as an attention clue. Patterns of 0, 17, 38, 60, or 82 dots composed the presented stimuli; with an exception of the zero, the number sequence is a geometric series on a logarithmic scale. We used stimuli, consisting of dots and displayed in random positions in order to prevent the constitution of patterns that may have otherwise influenced the results. Random patterns are usually considered as preferable to other configurations because the perceptual structures of the dot patterns could affect their apparent visual numbers (Frith & Frith, 1972; Ginsburg, 1976; Krueger, 1972). We decided to use a circular area with a fixed radius in order to prevent the number of dots from being proportional to the occupied area. A similar configuration has been widely used in many other experiments on this topic such as the studies by Knops, Viarougue, & Dehaene (2009) and Piazza, Izard, Pinel, Le Bihan, & Dehaene, (2004).

Procedure and Design. Participants were required to rate the numerosity of the presented dot patterns on a 0–20 numerical scale

(Anderson, 1962). Participants were instructed to consider the response scale with *none* (zero) and *very many* (20) as scale ends. Participants were also instructed to type the numerical scale point value that they rated on a keyboard. Each subject was shown three blocks: the training block and the sub-design block; for which the subjects were asked to rate the numerosity of sets of dots; and the factorial design block, for which the subjects were asked to rate the average numerosity between two sequentially presented dots sets (the experimental procedure is depicted in Figure 1). Participants were instructed to respond as quickly and accurately as possible and to not to try to count the dots.

Each trial was composed of a presentation part and a production part. In each presentation part, a circular gray area was shown at the center of the screen for 1000 ms, followed by the presentation of a dot pattern for 2000 ms. This gray area/dot pattern sequence was repeated twice. At the end of the presentation part of the trial, a hash mark (#) was presented for 1000 ms. The disappearance of the hash mark indicated the beginning of the response phase, in which the participants could type their responses. Participants typed their responses on a field on the screen by typing on a keyboard. After the participants made their judgments, they pressed a button to move on to the next trial, which started after an inter-trial interval of 500 ms.

Two subjects were excluded from data analyses because they did not show any response consistency.

Training block. Eleven trials were administered in order to familiarize the participants with the specific task and response method before the experimental blocks were given. Unlike the experimental blocks, in the training blocks, only one quantity per trial was shown and feedback for the participants was provided after each trial. The training block provided stimuli with a number of dots that ranged from 0–100, which represents the two anchors of the scale (Anderson, 1982). As a form of feedback, the computer provided the closest value on the 0–20 scale to the number of shown dots, divided by 5. This training allowed for the calibration of the judgments of numerosity and minimized the variability, caused by inter-individual differences in the perception of non-symbolic numerosity (see Izard & Dehaene, 2008).

Sub-designs block. Two dot patterns were presented. The participants were asked to rate the numerosity of one of them, either the first or the second, as indicated by a signal (number 1 or 2), presented after the disappearance of the second dot pattern. Each pattern could have one of five different numbers of dots: 0, 17, 38, 60, and 82. This five-by-five design yielded 25 pattern pairs. However, because no judgment different from 0 is plausible or informative, in response to “an empty” screen as a stimulus, target patterns with 0 dots were omitted; accordingly, only 20 (i.e., 4x5) pattern pairs were presented. Each pattern pair was presented twice, and each time, the pattern pair was presented with a different indication of the pattern to rate (number 1 or 2) for a total of 40 trials, presented in randomized order. In summary, we collected five responses for each dot pattern to be evaluated, and we used the mean of the five responses in the following statistical analysis.

Factorial design block. Participants had to rate the average quantity of dots between two presented patterns. Each pattern could have one of five numbers of dots: 0, 17, 38, 60 and 82. This five-by-five design yielded 25 pattern pairs. However, because no judgment different from 0 is plausible for pattern pairs with 0 dots, the (0, 0) pair was not presented; accordingly, only 24 (i.e., 5x5-1) pattern pairs were presented. Each pattern pair was presented 5 times for a total of 120 trials, presented in randomized order. In summary, we collected five responses for each pair of dot patterns to be evaluated, and we used the mean of the five responses in the following statistical analysis.

Each complete session of the experiment lasted approximately 30 minutes. Before every block, instructions were printed on the screen. Participants were requested to read the instructions and explain them back to the experimenter to verify that they understood correctly.

RESULTS

Psychophysical function. The shape of the response function of the sub-design block was tested. The shape of the response function using a magnitude estimation response methods is generally best described by a power function, $R = \alpha \times n^{\beta}$ with an exponent β smaller than 1 (Izard & Dehaene, 2008; see also Siegler & Opfer, 2003). In order to test the shape of the response function of the numerosity production response method, we performed a logarithmic regression analysis (see for instance Seber & Wild, 2003) for the estimations of each numerosity, averaged across subjects. Remarkably, the regression of the averaged data fits very well in $r^2 = .76$, and the resulting response function was $y = .93 \times n^{.73}$.

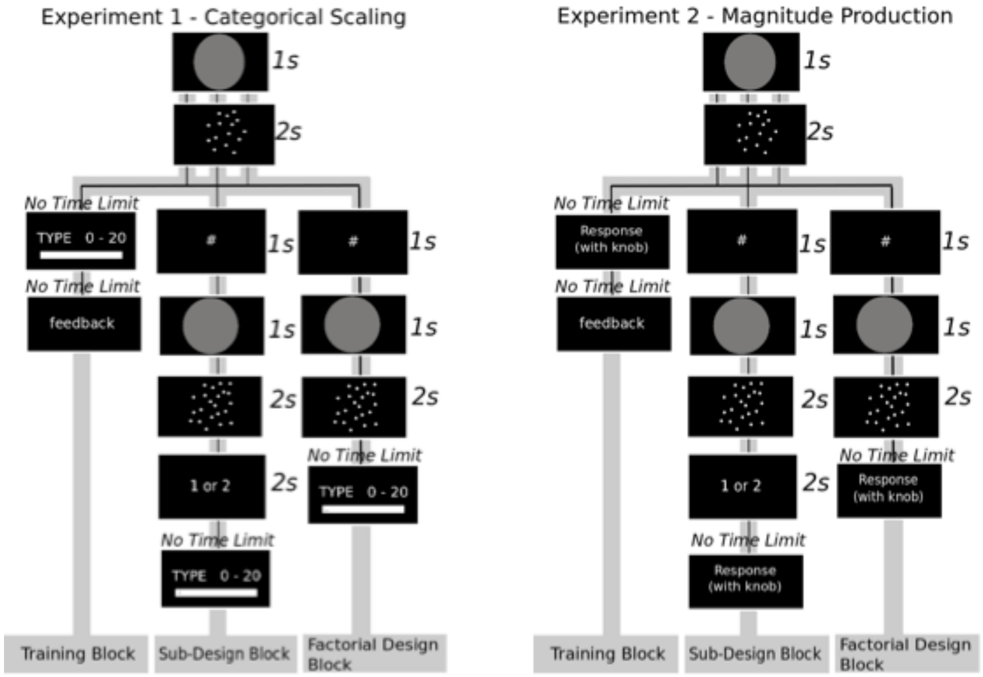

Figure 1. Experimental procedure for Experiment 1 (left panel) and Experiment 2 (right panel). The timing and sequence of events for the three blocks in each experiment are reported.

Model identification. The responses were analyzed in order to assess the plausibility of integration models. The classic approach, used by the functional measurement for the individuation of the integration function of the model, is the analysis of variance (ANOVA). The theorem of parallelism (Anderson, 1981) argues that if the integration model is additive, the graph of marginal means will appear as a bundle of parallel lines. Morever, any observed deviation from the parallelism will be purely due to the component of error. Thus, an ANOVA was conducted. Because of the interaction between the two factors (*1st and 2nd dot pattern*) was not significant ($F(15,14) = 1.52, p = .08$), the deviation from parallelism can be considered negligible, and the multiplicative model can be discarded from the candidates (see Figure 2). Moreover, a significant main effect was found for both factors: *1st dot pattern* ($F(4,14) = 198.45, p < .001, \eta^2_p = .309$) and

2^{nd} dot pattern (F(4,14) = 254.52, p < .001, η^2_p = .36). The test of the opposite effects (Anderson, 1981) is used to distinguish an additive model from an averaging one. This test makes use of the methodology of the sub-designs (Norman, 1976; Anderson, 1982). This methodology consists of associating the full factorial design with one or more sub-design(s) that exclude(s) one or more factor(s) at a time; the first experimental block was created explicitly for this purpose. The two factors (*1^{st} and 2^{nd} dot pattern*) were modified, adding to each one a level based on the responses of the sub-designs, referred to in the *1^{st} and 2^{nd} dot pattern*. If the model was not additive but averaging, then a significant interaction of the two factors was expected. Indeed, the ANOVA showed a significant interaction effect (F(15,14) = 2.30, p < .001, η^2_p= .014). Then, the parallelism observed in the full factorial design, along with the significance of the interaction, obtained when the sub-designs were added, might be considered as evidence in favor of the averaging model with equal weights within factors. It is the so-called equal-weight averaging model (EAM)(Wang & Yang, 1998). Moreover, for every factor, the significance of the main effect was found to be practically unaffected by the introduction of the new level, *1^{st} dot pattern* (F(4,14) = 199.00, p < .001, η^2_p= .251), *2^{nd} dot pattern* (F(4,14) = 249.52, p < .001, η^2_p= .296).

Model estimation. After the model was identified, the averaging model parameters for each participant were estimated with the R-average method (Vidotto & Vicentini, 2007; Vidotto, Massidda, & Noventa, 2010) and the implemented R-average package, version 0.4-0. The following analyses were computed on the estimated model parameters of all the participants, except when noted. The adaptation of the models to the data was evaluated, in terms of the adjusted r^2 for each subject, showing that the model fit the data very well for all of the participants of Experiment 1 with median r^2_{adj} = .84 (ranging between .78 and .99). As previously mentioned, the differential-weights model (DAM) was rejected, due to the lack of significant effects in the interaction between the linear components of the factors (Anderson, 1982). The EAM weights of the 1st and 2nd dot patterns were compared[1], revealing no significant difference (t(14) = -.60, p = .21).

[1] The maximum level of uniqueness (for w) is a common ratio scale. The unit of this scale is arbitrary because all the weights may be multiplied by a constant without changing the model prediction" (Anderson 1982, p. 89). Now, considering log(w), the origin of scale is arbitrary but no more the unit, indeed all the log(w) may be added by a constant with no change in the model prediction (Vidotto, 2013). In such a way the mean of log(w) has the property to be reference invariant and the standard deviation of log(w) has the property to

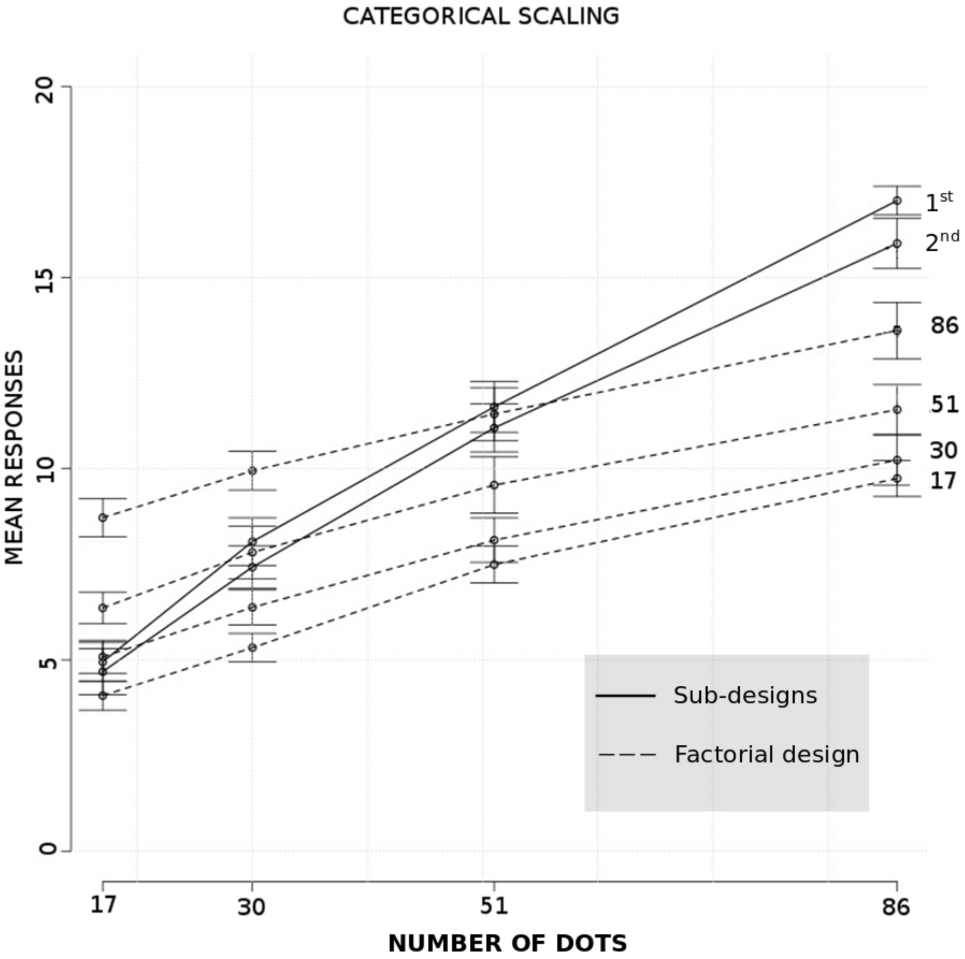

Figure 2: Experiment 1: plot of the subjects' estimations using categorical scaling method (mean responses are on the y-axis). In the complete factorial design, the number of dots identifies dashed lines for different numerosities of the 1st patterns while numbers of dots for the 2st pattern are in the x-axis. In the two sub-designs, 1st and 2nd identify continuous lines for the 1st and 2nd dot patterns while numbers of dots for the pattern are in the x-axis.

be absolutely invariant for any vertical translation; indeed, the t-test for differences was applied to log(w).

Under a principle of parsimony, this notion led us to opt for an averaging model with equal weights between factors ($w_A = w_B$), which we called a simple averaging model (SAM). A generalized linear mixed model (GLMM) was then applied to the *s* parameters of the SAM model, using the participants as random variables and the two factors, *numerosity* (0, 17, 38, 60, 82) and *dot pattern* (*1st, 2nd*), as fixed variables. The results showed a significant effect of the factor *numerosity* ($\chi^2(4) = 1786.61, p < .001$) with a strong and significant linear component. We found no statistically significant difference in the main effect for the dot pattern ($\chi^2(1) = 2.82, p = .27$) or interaction between numerosity and dot pattern ($\chi^2(4) = 3.66$, $p = .17$), showing that the difference between the two dot patterns in the *s* parameters was negligible.

Response latencies. The average latency to perform a categorical scaling was 4177.63 ms with a standard deviation of 1969.01 ms. It is important to note that the latencies were not correlated with the number of dots ($r = .04$). This result ensures that the participants were not using counting strategies; otherwise, we would have expected an increase in the reaction time with increasing numerosity (Akin & Chase, 1978; Mandler & Shebo, 1982; Trick & Pylyshyn, 1993).

EXPERIMENT 2: NUMEROSITY PRODUCTION

The aim of Experiment 2 was to test the appropriateness of a new method of numerosity production to IIT studies. We asked participants to indicate the numerosity of one presented dot pattern or the average numerosity between two sequentially presented dot patterns by producing that number of dots on the screen. Participants controlled the number of dots of their responses by turning a knob in a clockwise or counter-clockwise direction. To test the appropriateness of the method, we varied the number of presented dots systematically in a factorial design and sub-design. As in the previous experiment, we also studied the integration rule and evaluated the goodness of fit of the averaging model to the data. If participants responded on a linearly distributed scale, we would expect to find the plot of the complete factorial design in the shape of a bundle with parallel lines.

METHOD

Participants. Fourteen undergraduate female students from the University of Padua participated in the experiment. The average age of participants was 20.2 years (*SD* = .5). A convenience sampling was used. The participants received no payment.

Apparatus. The apparatus was identical to that in Experiment 1, except for the response device. The response device was a custom-made knob of 4.50 cm in diameter and 1.50 cm in height. The response device was also mounted on a small box (6 cm × 15 cm × 15 cm) and placed on a table. The knob was connected to a computer with a USB interface and could be rotated both clockwise and counter-clockwise. Knob rotation axis was parallel to the Cartesian *z*-axis. A Python program was developed to process the knob input and to control the stimulus presentation. The more the knob was rotated in a clockwise direction, the greater the number of dots that appeared on the screen. Rotation in the opposite direction decreased the number of dots, until no dots were left on the screen. Importantly, the spatial pattern of the appearance or disappearance of the dots was unpredictable. With every .6 degree of clockwise rotation, one additional dot (2 mm in diameter) was presented at a randomly chosen free position within an unmarked circular target area of 70 mm in diameter, centered on the screen. Counter-clockwise rotation deleted randomly selected dots from the display. The minimum distance between the two dots was .25 mm. The maximum number of dots was limited to 300.

Materials. The materials used were the same as those in Experiment 1.

Procedure and Design. The procedure and design were identical to those of Experiment 1, except for the response method that consisted of rotating the knob to produce the desired quantity of randomly distributed white dots (see Figure 1). Thus, the response method in this experiment was a numerosity production, instead of a categorical scaling. At the beginning of the response phase, participants could rotate the knob in order to perform the numerosity production task. Participants always started the response phase with zero dots on the screen and turned the knob clockwise to increase the number of dots or counter-clockwise to decrease it.

RESULTS

Psychophysical function. As for the data of the sub-design in Experiment 1, in order to test the shape of the response function of the categorical scaling response method, we performed a logarithmic regression analysis for the estimation of each numerosity, averaged across subjects. The regression of the averaged data fitted acceptably with $r^2 = .58$. The resulting response function was, $y = .15 \times \mathrm{n}^{.58}$.

Model identification. In Experiment 1, the responses were analyzed in order to assess the plausibility of integration models by performing a 5x5 ANOVA with *1st and 2nd dot pattern* as factors. Because the interaction between the two factors was not significant ($F(15,110) = .82$, $p = .64$), the deviation from parallelism can be considered negligible, and the multiplicative model can be discarded for this experiment, as it was for Experiment 1. Moreover, a significant main effect was found for every factor, *1st dot pattern* ($F(4,11) = 162.35$, $p < .001$, $\eta^2_{\mathrm{p}} = .28$), *2nd dot pattern* ($F(4,11) = 227.90$, $p < .001$, $\eta^2_{\mathrm{p}} = .358$). The two variables (*1st and 2nd dot pattern*) were modified, adding to each one a level that was made from the ratings on single-dot patterns (the sub-design). If the model was not additive, but averaging, it was expected that the addition of the new levels to the factors would involve a significant interaction between the two. Subsequently, the ANOVA showed a significant interaction effect ($F(15,110) = 2.21$, $p < .001$, $\eta^2_{\mathrm{p}} = .01$). Then, the parallelism between the factors of the full factorial design and the interaction, caused by the adding of sub-designs (see Figure 3) was found to be evidence in favor of EAM, as in Experiment 1. Moreover, for each factor, the significance of the main effect was found to be practically unaffected by the introduction of the new level, *1st dot pattern* ($F(4,11) = 172.12$, $p < .001$, $\eta^2_{\mathrm{p}} = .241$), *2nd dot pattern* ($F(4,11) = 229.78$, $p < .001$, $\eta^2_{\mathrm{p}} = .298$).

Model estimation. After the model was identified, the averaging model parameters for each participant were estimated with the same procedure that was previously applied in Experiment 1. The adaptation of the models to the data was evaluated in terms of adjusted r^2, showing that the model fitted the data very well for almost all of the participants with median $r^2_{\mathrm{adj}} = .85$ (ranging between 71 and 99). As previously mentioned, the DAM was rejected because it did not present a significant effect in the interaction between the linear components of the factors. The EAM weights of the first and the second dot patterns were compared, which revealed no significant difference ($t(11) = 1.79$, $p = .56$). This led us to opt for a SAM.

A GLMM was then applied to the *s* parameters of the model, using the participants as random variables and the two factors of *numerosity* (0, 17, 38, 60, 82) and dot pattern (*1st, 2nd*), as fixed variables. We found that there is a statistically significant difference in the main effect for *numerosity* ($\chi^2(4) = 4044.78$, $p < .001$) but not for the dot pattern ($\chi^2(1) = 1.50$, p=.13) or for the interaction between *numerosity* and dot pattern ($\chi^2(4) = 5.02$, $p =$ .10). This shows that the difference between the two dot patterns in the *s* parameters was negligible.

Response latencies. The average latency to perform a categorical scaling was 3880.13 ms (*SD* = 2229.38). Importantly, the latencies were not correlated with the number of dots (*r* = .039), ensuring that the participants were not using counting strategies, which was also the case in Experiment 1.

DISCUSSION AND CONCLUSIONS

In both experiments, the participants responded quickly, and their response times did not increase with numerosity. This reveals that the participants did not use counting strategies but instead, based their judgments on approximate numerosity estimation. In both of the experiments, the results of the analysis on the estimated averaging values seemed to indicate that the subjects' estimations are best described by an EAM. Moreover, the weights of the two dot patterns do not appear to differ significantly, suggesting the use of a SAM. Accordingly, the scale values vary, depending only on the numerosity of the stimulus and are unaffected by its position (*1st or 2nd dot pattern*). This demonstrates that neither the effect of primacy nor the effect of recency influence the evaluation of the average numerosity, despite the sequential temporal order of the presentation of the stimuli (Busemeyer, 1991). In other words, this means that the participants give the same importance to the two quantities of each trial during averaging operations.

In both experiments, the adjusted r^2 showed that SAM was able to explain a very great portion of variance for almost all of the participants; this supports the explanatory capability of the averaging model, applied to mental arithmetic problems with discrete quantities. Since the participants were instructed to perform an averaging operation, the factorial plot should exhibit parallelism, if the response measure was on a linear scale. As shown in Figures 2 and 3, and according to the results of the full factorial design ANOVA (without sub-designs), the rating data (Figure 2) and the numerosity production data (Figure 3) show clear parallelism. This allows researchers to validate the numerosity production, as a response measure on a linear scale, a prerequisite for a method to study stimulus interaction, and for the analysis of non-linear integration rules (Anderson, 1982).

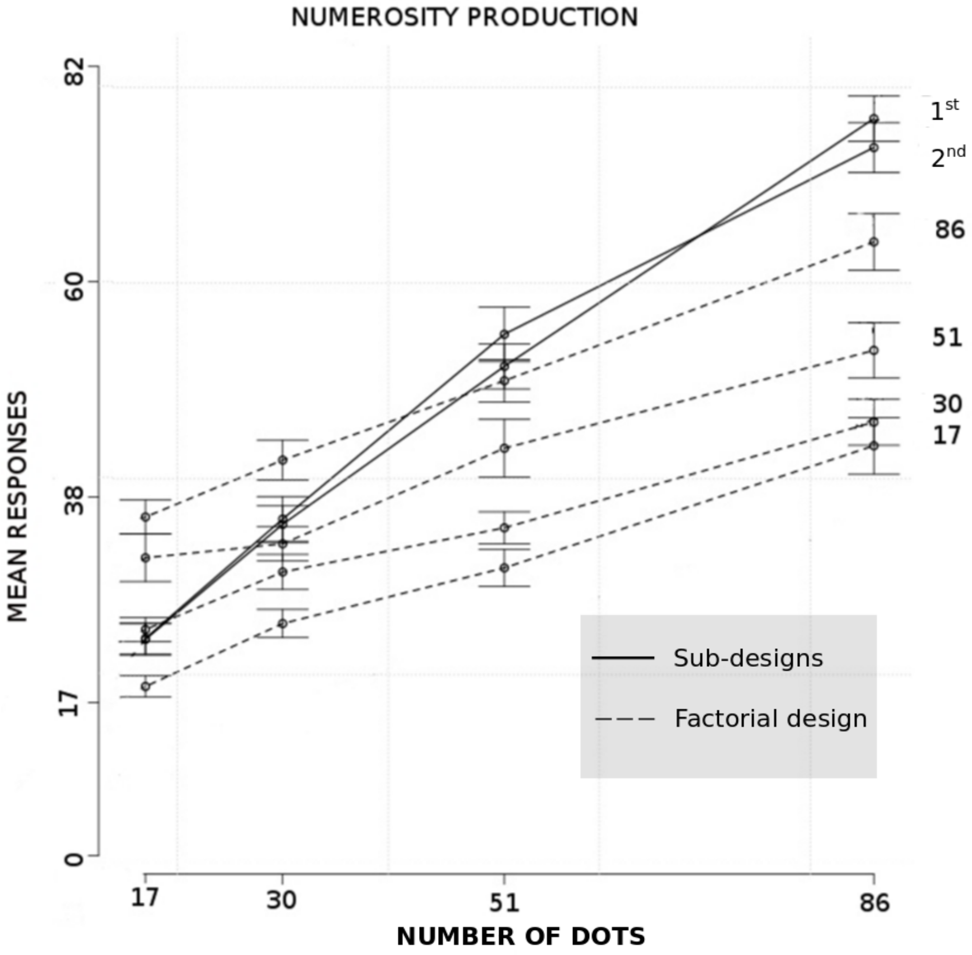

Figure 3: Experiment 2: plot of the subjects' estimations using numerosity production method (mean responses are on the y-axis). In the complete factorial design, the number of dots identifies dashed lines for different numerosities of the1st pattern while the numbers of dots for the 2st pattern are in the x-axis. In the two sub-designs, 1st and 2nd identify continuous lines for the 1st and 2nd dot patterns while the numbers of dots for the pattern are in the x-axis.

The linearity of the response scale and the similar trends of the two response methods paves the way for further interesting possibilities of application of the IIT framework for the numerosity production response method. The applicability of IIT to mental arithmetic problems with discrete quantities is supported with the following factors: the linearity of the scale observed with both of the response methods and the high explanatory capability demonstrated by the averaging model in both experiments.

On the other hand, since we used a series of stimuli, composed by dot collectivities, distributed on a fixed radius circular area and manipulated the number of dots, it may be argued that the density of dots in each stimulus may have influenced the participants' impressions of numerosity (Krueger, 1972; Allik & Tuulmets, 1991; Shanteau et al., 2007). We believe that the variation in density, along with the levels of the factors, do not weaken our conclusions. This is because even if the numerosity judgment was based on the density of the stimulus, it does not change the way that the internal representations of the stimuli were integrated. Furthermore, this does not change the conclusions about the parallelism and linearity of the response functions. Since the effect of over- or under-evaluation, linked to the specific density of each level of each factor is proportional to the size of the stimulus, and since it remains constant for that level to every proposition in the factorial design, this does not affect the nature of the model but affects only its scale values.

The averaging model of IIT was established as a viable instrument in assessing mental arithmetic with discrete quantities; it is able to properly describe behavioral data, distinguishing between the value of the evaluation of a stimulus and its importance in the integration process. Moreover, a new numerosity production method was tested for the linearity of its response scale. Finally, averaging operations with discrete quantities appear to not be affected by the presentation order of the dot patterns. For all of these reasons, the IIT framework seems to be a promising approach, particularly for future applications in the field of mental arithmetic with discrete quantities.

REFERENCES

Akin, O., & Chase, W. (1978). Quantification of three-dimensional structures. *Journal of Experimental Psychology: Human Perception and Performance, 4,* 397-410.

Allik, J. (1989). Is a unified psychophysical law realistic. *Behavioral and Brain Sciences*, 12(2), 267-268.

Allik, J., & Tuulmets, T. (1991). Occupancy model of perceived numerosity. *Perception & Psychophysics, 49*(4), 303-314.

Anderson, N. H. (1962). Application of an additive model to impression formation. *Science, New Series*, *138*(3542), 817-818.

Anderson, N. H. (1965). Averaging versus adding as a stimulus combination rule in impression formation. *Journal of Experimental Psychology*, *70*(4), 394-400.

Anderson, N. H. (1981). *Foundation of information integration theory*. New York: Academic Press.

Anderson, N. H. (1982). *Methods of information integration theory*. New York: Academic Press.

Anderson, N. H., & Shanteau, J. C. (1970). Information integration in risky decision making. *Journal of Experimental Psychology*, *84*(3), 441-451.

Anderson, N. H., & Weiss, D. J. (1971). Test of a multiplying model for estimated area of rectangles. *American Journal of Psychology, 84*, 543-548.

Binet, A. (1969). Children's perceptions. In R. H. Pollack & M. J. Brenner (Trans., Eds.), *The experimental psychology of Alfred Binet* (pp. 93-126). New York: Springer. (Original work published 1890).

Busemeyer, J. R. (1991). Intuitive statistical estimation. In N. H. Anderson (Ed.), *Contributions to information integration theory. Volume 1: Cognition.* Hillsdale, New Jersey: Lawrence Erlbaum.

Cuneo, D. O. (1982). Children's judgments of numerical quantity: A new view of early quantification. *Cognitive Psychology, 14*, 13-44

Dehaene, S. (1997). *The number sense: How the mind creates mathematics*. New York: Oxford University Press.

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

De Marchi, S. (1986). Numerical evaluations of collectivities. (S. C. Masin, Trans.). (Original work published 1890). Retrieved from http://www.psy.unipd.it/~masin/DeMarchi1929.pdf

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Science, 8*(7), 307–314.

Frith, C.D., & Frith, U. (1972). The Solitaire Illusion: An illusion of numerosity. *Perception and Psychophysics, 11*, 409-410.

Ginsburg, N. (1976). Effect of item arrangement on perceived numerosity: Randomness vs regularity. *Perceptual & Motor skills, 43,* 663-668.

Izard, V., & Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, *106*(3), 1221-1247.

Knops, A., Viarouge A., & Dehaene, S. (2009). Dynamic representations underlying symbolic and non-symbolic calculation: Evidence from the operational momentum effect. *Attention, Perception & Psychophysics, 71(4)*, 803-821.

Krueger, L. E. (1972). Perceived numerosity. *Perception & Psychophysics, 11,* 5-9.

Lourenco, S. F., & Longo, M. R. (2011). Origins and the development of generalized magnitude representation. In S. Dehaene & E. Brannon (Eds.), *Space, time, and number in the Brain: Searching for the foundations of mathematical thought* (pp. 225-244). London, England: Elsevier.

Mandler, G., & Shebo, B. (1982). Subitizing: An analysis of its component process. *Journal of Experimental Psychology: General, 111*, 1-22.

Norman, K. L. (1976). A solution for weights and scale values in functional measurement. *Psychological Review*, *83*(1), 80-84.

Piazza, M., Izard, V., Pinel, P., LeBihan, D., & Dehaene, S. (2004). Tuning curves for approximate numerosity in the human parietal cortex. *Neuron, 44(3)*, 547-555.

Shanteau, J., Pringle, L. R., & Andrews, J. A (2007). Why functional measurement is (still) better than conjoint measurement: judgment of numerosity by children and adolescents. *Teorie & Modelli, 12*, 199-210.

Seber, G. A. F., & Wild, C. J. (2003). *Nonlinear Regression*. New York: Wiley-Interscience.

Siegler, R. S., & Opfer, J.E. (2003). The development of numerical estimation: Evidence for multiple representations of numerical quantity. *Psychological Science, 14*(3), 237-243.

Trick, L., & Pylyshyn, Z. (1993). What enumeration studies can show us about spatial attention: Evidence for limited capacity preattentive processing. *Journal of Experimental Psychology: Human Perception and Performance, 19*, 331-351.

Vidotto G. (2013). Note on differential weight averaging models in functional measurement. *Quality & Quantity*, *47*(2), 811-816.

Vidotto, G., Massidda, D., & Noventa, S. (2010). Averaging models: Parameters estimation with the R-Average procedure. *Psicòlogica*, *31*(3), 461-475.

Vidotto, G., & Vicentini, M. (2007). A general method for parameter estimation of averaging models. *Teorie & Modelli*, *12*(1-2), 211-221.

Wang, M., & Yang, J. (1998). A multi-criterion experimental comparison of three multi-attribute weight measurement methods. *Journal of Multicriteria Decision Analysis*, *7*(6), 340-350.

Zalinski, J., & Anderson, N. H. (1989). Measurement of importance in multi-attribute models. In J. B. Sidowski (Ed.), *Conditioning, cognition, and methodology. Contemporary issues in experimental psychology* (pp. 177-215). Lanham, Mariland: University press of America.

Zalinski, J., & Anderson, N. H. (1991). Parameter estimation for averaging theory. In N. H. Anderson (Ed.), *Contributions to information integration theory. Volume I: Cognition* (pp. 353-394). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

(Manuscript received: 21 November 2013; accepted: 22 June 2014)