Making Numbers Come to Life:

Two Scoring Methods for Creativity in Aurora's Cartoon Numbers

Mei Tan[1*]

Catalina Mourgues[2]

David S. Bolden[3]

Robert J. Sternberg[4]

Elena L. Grigorenko[1,5,6]

[1]Yale University, USA
[2]Pontificia Universidad Catolica de Chile, Chile
[3]Durham University, UK
[4]Oklahoma State University, USA
[5]Columbia University, USA
[6]Moscow State University, Russia

*Corresponding Author
Mei Tan
Yale University
230 South Frontage Road
New Haven, CT 06519-1124
Phone: 203.785.4831
Fax: 203.785.3002

Abstract

While creativity has long been recognized as an important aspect of mathematical thinking, both for the advancement of the field as well as in students' developing expertise in mathematics, assessments of student creativity in that domain have been limited in number and focus. This article presents an assessment developed for creativity that provides a score for mathematical creativity (MaC) in addition to a score for general creativity in the numeric domain, or what we might call numerical creativity, (NuC). We developed different rating scales for each and then explored how each scoring method accounts for the students' mathematical/numerical and creative skills. The psychometric properties for both scoring approaches were examined. Each method was shown to reflect different relationships with other performance tests. In addition, it is proposed that MaC may provide useful insight into students' level of adaptive expertise in mathematics, as reflected by their ability to apply mathematical knowledge (i.e., language, operations, concepts) to novel situations, representing an informative supplement to performance indicators of math achievement.

Keywords: creativity, numeracy, mathematics, assessment

The field of creativity research has come a long way since the pioneering efforts in the field of J. P. Guilford (1950) and E. Paul Torrance (1962, 1974) in the 1950s and 60s (Sternberg, 2006). Research in the field has grown to investigate creativity at all levels of expertise and in a wide-range of activities in various domains. Yet, many questions concerning that construct remain to be answered, and recurring themes include the definition of creativity itself. In this article, building on the basic, generally accepted definition of creativity as the ability to produce original work within particular task constraints (Lubart & Guignard, 2004; Lubart, Pacteau, Jacquet, & Caroff, 2010; Lubart & Sternberg, 1994; Sternberg, Lubart, Kaufman, & Pretz, 2005), we explore the manifestations of creativity in the domains of number knowledge and mathematics through two different scorings of a single sample who were administered a new assessment for creativity.

The consideration of the role of creativity in mathematics began with mathematicians who pondered the nature of the cognitive processes that contributed to mathematical discovery, what they called the psychology of invention (Hadarmard, 1945; Poincaré, 1913), which seemed to proceed in stages of preparation, incubation, and illumination. Creativity has since been identified as necessary for the advancement of the field of mathematics, along with being an important aspect of mathematical thinking in students (e.g., Bolden, 2010; Krutetskii, 1976; Mann, 2006; Polya, 1962; Sriraman, 2004; Sternberg, 1996). But what are the manifestations of creativity in mathematics as it is developing below the level of expert or genius, in the realm of the student? Over the last 30 years or so, educators and researchers have attempted to define what constitutes mathematical creativity in students, coming up with the following: the ability to see new relationships and make associations between seemingly unrelated ideas (Haylock, 1997); the ability to engage in problem formation and invention with independence and originality (Krutetskii, 1976); the ability to overcome functional fixedness, to think beyond given or conventional mindsets (Balka, 1974; Haylock, 1987, 1997; Krutetskii, 1976); and the ability to abstract and generalize mathematical material (Krutetskii, 1976; Sternberg, 1985). These definitions for the most part are derived from the application of general creativity within the domain of mathematics, across a variety of mathematical activities.

Yet, the role of creative thinking in less abstract aspects of mathematics, such as problem-solving and problem-finding, has been discussed and investigated more extensively (Krutetskii, 1976; Pehkonen, 1995; Polya, 1962). These discussions have formed the basis of most assessments of creativity in mathematics. Getzels and Jackson’s study of types of giftedness in children (1962) provided the model for many later assessments for creative potential with numbers. In their “Make-up Problems” test, children read complex paragraphs rich in numerical information and were asked to create as many mathematical problems as they could with the given information. Scores were calculated to reflect the number, appropriateness, complexity, and originality of the problems created, generally mirroring the classical evaluation of creativity in other domains. Each problem was scored for the numerical information and operations it contained.

Subsequent measurements have focused similarly on creative products with respect to the amount produced (fluency), the range of types of what has been produced (flexibility), and the relative novelty of productions with respect to what has been produced by others (originality) (Balka, 1974; Haylock, 1997; Kim, Cho, & Ahn, 2003). For example, Balka (1974) provided students with mathematical situations from which to develop problems; creativity was assessed on the flexibility, fluency, and originality of the problems produced. Problem solving of open-ended questions is another main method of fostering creativity (Pehkonen, 1995, 1997); methods for promoting creativity through mathematical problem solving are presented by Hashimoto (1997). Cho and colleagues developed an assessment for mathematical creative problem-solving that looked at the number of valid solutions students could produce for open-ended math questions (Math Creative Problem Solving Ability Test, MCPSAT; Kim et al., 2003). These open-ended questions were designed to require math skills and content knowledge, convergent and divergent thinking, and task commitment; they were scored for fluency (Cho & Hwang, 2006). Leikin and Lev (2007) elicited multiple solutions to both conventional and unconventional math problems to compare the performance of mathematically gifted, mathematically proficient, and typically performing math students. These were scored for novelty, flexibility and fluency, and it was found that unconventional math problems may be more useful for differentiating students who are mathematically creative and, most likely, mathematically gifted, although it should be noted that some suggest that it is possible to be mathematically gifted without being mathematically creative, and vice versa (Sriraman, 2005). Haylock (1987, 1997) proposes assessments for mathematical creativity that focus on the ability to overcome fixations in the application of mathematical content (i.e., sums and differences) and procedures (algorithms), along with the ability to formulate many responses to open-ended questions (i.e., requests for problem-solving, problem-posing, redefinitions).

The study here presents data from a new assessment for creativity, Cartoon Numbers (CN), one of the subtests of the Aurora Battery, a cognitive ability battery being developed for children aged 9-12, based on Robert Sternberg’s theory of successful intelligence (for more on Aurora and successful intelligence, see Chart, Grigorenko, & Sternberg, 2008; Kornilov, Tan, Elliott, Sternberg, & Grigorenko, 2011; Sternberg, 1999, 2010). CN was designed to assess creativity in the numeric domain, and it therefore differs from those assessments for mathematical creativity described above in that it is not directly about mathematics; no equations are presented for solving, and the scenarios do not provide quantities to be used in any suggested calculations. Rather, students are presented with social scenarios in which the characters are numbers, and then are asked to explain the situation or propose a solution to a social conflict. That is, CN's open-ended nature leaves students the option to respond in a narrative and/or mathematical way. What became interesting in the course of scoring these tests were the number of children who chose to explain these social situations primarily or solely in terms of mathematical equations or concepts. The original rating scale for scoring CN focuses on the exercise of creativity within the domain of numbers ("numeric creativity," NuC), defining this as the ability to integrate number knowledge (by imagining or integrating number concepts or shape) into non-numerical or novel contexts, i.e., how well numbers can be imagined as social characters, or how well a social situation may be explained using numeric knowledge (Barwell, 2004). According to this rating scheme, each item is scored for the most creative elements in the given response. Narrative responses that combine social and mathematical descriptions are considered appropriate, while a response containing math only is considered incomplete. The most creative responses are those that elaborately and originally present number behavior in a social context. The NuC rating scale does not consider the quantity of responses (fluency) as an index of creativity, although the children may give more than one good answer.

Alternatively, the new rating scale proposes a CN score for creative mathematical thinking (MaC). Rather than focusing on children's ability to imagine numbers as social and emotional characters, this rating scale examines children's ability to model the given situation in terms of mathematics in unique and/or multiple ways. This may be defined as creatively applied mathematical ability, or the use of mathematical language, concepts, relationships or operations to logically structure or explain a novel situation (Aiken, 1973; Krutetskii, 1969).The significance of this new rating scale is its focus on the use of mathematics for the purpose of analogy or as causal metaphor, to be applied within a context (i.e., social) that is outside of the mathematical domain. The modeling of math onto non-mathematical situations was not considered in the original rating scale and is also an aspect of mathematical creativity not typically emphasized in mathematical creativity assessments thus far. According to the MaC rating scale, all mathematical equations, operations and concepts that appeared to logically explain the given social situation were considered appropriate. The quantity of these types of responses were taken as indicators of fluency; originality was the relative rarity of the concepts applied within the cohort. Yet, as noted by Runco, Okuda and Thurston (1987), simply summing the fluency and originality scores for a total creativity score incorporates information that is redundant, since the fluency score takes into account both original and non-original responses. In addition, greater fluency increases one's chances of producing something original (Mouchiroud & Lubart, 2001; Runco et al., 1987), thus providing a kind of advantage for originality. To counterbalance the advantages of fluency, we tested four methods for the calculation of a total mathematical creativity score that would include aspects of both originality and fluency but in a weighted manner. Flexibility was not considered in our scoring because the variety of mathematical concepts that could be generated was limited by the nature of the items.

Below, we present the psychometric properties of both the NuC and MaC rating scales for the CN subtest, and examine how each scoring method accounts for the students' mathematical and creative skills. To explore CN's possible relationship to academic achievement and related skills, we compared both scorings to students' corresponding Key Stage 2 (KS2) or Middle Years Information Systems (MidYIS) tests. We hypothesized that the two rating scales would provide distinct sets of scores and show different patterns of relationships with other types of tests. The MaC rating scale, we posited, would exhibit a closer connection to academically-oriented assessments, yet not necessarily to math assessments, as MaC measures an aspect of mathematics skill that is not generally examined in traditional tests.

Method

Participants

The sample for this study was 205 children; 52.9% were female (109) and 47.1% were male (96); 22.3% were in 4th grade, 43.2% in 5th grade, and 34.5% in 6th grade. The average age for this group was 10.39 years (*SD* = 1.37 years, min = 8.41 years and max = 13.08 years). The participants were drawn from a larger sample of 2,818 British school students, recruited through schools located in two towns in England. Twenty-six schools were selected as representative of the larger population (8 secondary schools and 18 primary schools), and 205 children were randomly selected to represent a range of grades and schools. This population is largely homogeneous and is from one of the most disadvantaged areas in the UK.

Measures

Cartoon Numbers. Cartoon Numbers (CN) is a subtest of the Aurora Battery. Aurora was developed to assess analytical, creative, and practical abilities in a group or classroom setting. The battery is characterized by variation in its item formats (multiple choice, short answer and open-ended items, scored by trained raters). The subtests were designed to assess the above abilities across and between stimulus domains (verbal, numerical, and figural subtests) and item formats such that a balanced range of opportunities could be offered for children to demonstrate various abilities within and across domains (Chart et al., 2008). CN elicits open-ended responses and was designed to target creative ability in the numerical domain.

CN has seven items, each presenting a simple social situation that would be expected to be familiar to children, in which the characters are depicted by numbers. No item involves more than three characters. The directions read: “Here are some interesting questions using cartoon numbers. Read the questions and use your imagination to answer them in a creative way. There are no wrong answers!” The example given for the subtest is: “Number 2 and number 4 are talking, and they are having a really good time. Why are 2 and 4 getting along so well?” An example of what would be considered a less creative response to this question might be “They just met”. The following example response would be considered more creative: “2 and 4 get along so well because they have a lot in common! They are both even numbers. 2 is happy that when he is added to another 2, they are like 4. And 4 thinks it’s great that when 2 and 4 get together they make 24, the number of hours in a day!” CN is clearly not purely mathematical in nature, but was designed to be a more general creativity test with domain-specific elements. Its seven items were varied in content to provide more complex scenarios for a range of item difficulty, but were intended to be homogeneous in nature to draw upon the same abilities. Yet the mathematical aspects of the responses it ultimately elicited were notable and worthy of examination as they indicated a readiness to map mathematical thinking onto non-mathematical situations.

We examined two approaches for scoring the CN subtest. For the first approach (NuC), we used the following scoring rating scale for each item. One score was given for accuracy or task appropriateness (0-2) and one score for creative ability (0-4). Task appropriateness indicated whether or not the child followed the instructions, i.e., was the response appropriate to the question. According to this rule, responses lacking any social element or consisting only of mathematical equations were scored a 1 for task appropriateness, while a 0 was defined as completely off-task or nonsensical. Creative ability points were given based on the novelty, complexity and elaboration of the response, with the best responses giving numerical justifications for the behavior of the number characters, applying numerical concepts to social situations. Also, it was considered that the invention of a story or set of facts that reasonably explains the social behavior of numbers reflects students' creative ability to cope with uncertain or novel situations (Sternberg, 1999, 2006). Hence, the NuC rating scale recognizes the integration of social knowledge with numerical knowledge in unique and elaborate responses that could include number values or operations, as well as shape. Notably, responses were scored for their best elements—with respect to elaboration and originality—without recognizing the quantity of the responses provided, i.e., the fluency. A score of 0 for accuracy was given for off-task or nonsensical responses; a score of 4 was given for elaborate, original responses incorporating both numeric and social knowledge that appeared to address the question. Analyses were carried out using a total (summed) score that considered both accuracy and creative ability. This follows the generally accepted definition of creativity as the ability to produce original work (i.e., ability score) within particular task constraints (i.e., accuracy score) (Lubart & Guignard, 2004; Lubart et al., 2010; Lubart & Sternberg, 1994; Sternberg et al., 2005).

The second scoring approach (MaC) that we explored in this study assessed students' originality and fluency in their application of mathematical concepts to the social situation. Credit was given for math concepts only if they were applied appropriately (made sense); nonsensical math was scored as 0. For each item, the number and frequency of appropriately applied math concepts, as well as their frequency by type across items, were calculated for the total sample. The most infrequent concepts for each item were considered more original. The total number of mathematical concepts applied in each item produced a fluency score.

To obtain a total MaC score for each item, we tested four methods. The first—the *Weighted Originality Sum (WOS)* method, similar to that used by Runco, Okuda and Becky (1987)—calculated originality as the relative frequency of each math concept applied in each item of CN. The relative frequency was then subtracted from 1, so that less frequent concepts had a greater number while the most frequent mathematical concepts had a lower number. Finally the weights were summed. This method permitted the consideration of both the fluency and originality while reducing the redundancy of the scores (i.e., if you produce more responses, the more likely you are to produce one that is original). The second method was the *Ratio Weight Originality Sum (RWOS)*, using the same method as above, but this time controlled for fluency by dividing the WOS by the number of mathematical concepts given by the child. This ratio score was intended to take into account the different weights of the originality scores within items, adjusting for fluency. The third method was the *Ranking Ordinal Sum (ROS)*, which generated scores by calculating the rank of mathematical concepts within each item. A rank of 1 indicated mathematical concepts of the highest frequency; higher values were assigned to the concepts that appeared lowest frequently. This sorted the responses according to their originality without considering fluency. Finally, the *Ratio Ranking Ordinal Sum (RROS)* method was used, which controlled for fluency by dividing the final score (ROS) by the number of mathematical concepts given by the child.

Key Stage 2. The Key Stage tests are part of the English system of academic tracking; at the end of particular educational “stages,” all students are assessed (using various combinations of paper and pencil tests and teacher assessments) to determine whether or not they are working at the expected age-based level for each of several academic subjects. KS2 tests (which are re-written each year) are administered to students in the UK at the end of their Year 6 (grade 5);

they contain tests of Reading, Writing, English, Mathematics and Science[1]. The Mathematics section contains problems that are directly related to the UK National Curriculum topics of number, algebra, shape, space and data handling (DirectGov, 2009).

MidYIS We were also able to compare students' CN scores to their scores on a more consistent assessment, the MidYIS (Middle Years Information System; Centre for Evaluation and Monitoring, 2010). The MidYIS, designed to be taken by students upon entry to secondary school (during Term 1 of Year 7 or at the very end of Year 6), is described as a baseline assessment of developed ability and aptitude for learning. The four sections of the MidYIS are vocabulary, mathematics, skills (proofreading, perceptual speed, and accuracy), and nonverbal tasks (primarily visual-spatial reasoning and logical thinking). The mathematics section is designed to access speed and fluency, rather than mathematics knowledge (www.midyisproject.org). According to the Centre for Evaluation and Monitoring (CEM), who produce and manage the tests, vocabulary and mathematics scores in particular are predictive of student achievement (as reflected in students' scores on the standardized national achievement—Key Stage—tests). The MidYIS overall score, which we used to compare to Aurora's scores, is calculated by adding the weighted raw scores for the vocabulary, mathematics, nonverbal, and skills sections.

Procedure

Cartoon Numbers was administered to classes in school as part of Aurora-*a* (the Aurora Battery's triarchically-based paper and pencil test). Aurora-*a* was given to all classes in multiple sessions over the course of three days. For Year 7 (U.S. Grade 6) each daily session was one hour. The battery was split into 3 packets—A, B, and C, and the order of the subtests in the packets was counterbalanced across two versions, 1 and 2. For Years 6 and 5 (U.S. Grades 5 and

[1]Science was discontinued as a routinely administered Key Stage subject in 2010.

4) each daily session consisted of 2 sessions of 45 minutes each, with a 15-minute break between sessions. The test battery was thus split into 6 packets (A, B, C, D, E, and F). The order of subtests in each packet was counterbalanced in two versions, 1 and 2; the order of the test packets in all data collections was counterbalanced across schools. In this particular case, KS2 was administered to these students in May of 2006, and the MidYIS in the fall of 2006; Aurora was administered to the students in the January of the following year, 2007.

Data analysis

The reliability between the judges was calculated using Cronbach's α for each rating scale, to assess the internal consistency of both the NuC and MaC scoring methods. Based on the assumption that the items of CN are explained by a single latent variable—numerical or mathematical creativity—Confirmatory Factor Analyses (CFA) were conducted for one factor models for both the NuC and MaC scores. The CFAs were conducted using AMOS software (Version 17.0). Goodness-of-fit was determined by evaluating (1) the ratio of C-MIN to the degrees of freedom (C-MIN/*df*), with values of < 1.00 constituting a good fit and values between 1.00 and 2.00 reflecting acceptable fit; (2) the Root Mean Square Error of Approximation (RMSEA), with values < .06 constituting a good fit; and (3) Comparative Fit Index (CFI) values > .95 constituting a good fit and values >.90 acceptable fit (Byrne, 2010). In the case of an unacceptable fit of the model to our data, an Exploratory Factor Analyses (EFA) was performed using principal axis factoring analyses with varimax rotation. Two criteria were used to decide the number of factors: the Kaiser eigenvalue criteria greater than one and the scree plot. The KMO and Bartlett test were reported for each EFA.

To assess both rating scales' associations with Aurora-*a*, bivariate correlations were obtained for the factor scores of the MaC and NuC scales. Aurora's subtests were grouped according to their respective target abilities (analytical, creative, and practical) and domains (verbal, numerical, and figural), and their total ability/domain scores were calculated for the correlation analysis. Finally, to determine whether the NuC and MaC scores were related to mathematical achievement or other mathematical skills assessed with the MidYIS, two independent structural equation models were tested. Given the small sample size for the KS2 data, its relation with NuC and MaC rating scales was assessed using independent linear regressions for each subtest (these were controlled for gender in case of significant differences between boys and girls).

Results

Psychometric proprieties of the NuC and MaC rating scales

In this section, we compare the psychometric performance of the NuC and MaC rating scales, and the four alternate scoring methods for MaC. For the NuC rating scale, 12.9% of the tests were scored by two judges independently. The agreement between judges measured with Cronbach's α was .895; the internal consistency of the seven items was .610. For the MaC rating scale, 10% of the tests were scored by two judges independently; the agreement between judges measured with Cronbach's α was .98; the internal consistency for the seven items was .750 when considering the raw scores. For the MaC rating scale, using the *Weighted Originality Sum (WOS)* score, Cronbach's α was .742; for the *Ratio Weighted Originality Sum (RWOS)* score, .673; for the *Ranking Originality Sum (ROS)*, .652; and for the *Ratio Ranking Originality Sum (RROS)*, .418. This lack of reliability for the *RWOS* and *RROS* methods may have been due to the control of fluency, revealing that not all of the items had the same probability of generating the same quantity of responses. This item heterogeneity is discussed at length in the Discussion section. In addition, the *WOS* method reflects a more accurate scaling of originality than the *ROS* method by using the observed frequency of a response across the sample to represent originality, rather than ranking, which represents only relative originality. The following analyses will report on findings from the NuC rating scale and the MaC rating scale using the *WOS* method (from now on referred to as the MaC score), as this method appears to provide a more consistent method for evaluating mathematical creativity across all of the subtest's items.

The descriptive data for both rating scales are reported in Table 1.The means and standard deviations of the items for both rating scales were expressed as standardized scores for better comparison; the standardized scores were calculated for each rating scale independently using the mean and standard deviation for the complete test. Interestingly, the NuC scores show the highest averages for the last items, while the MaC scores show the highest averages for the first items. For example, for Item 1, the NuC average was -.12 below the mean, while the MaC average for the same item was .52 above the mean. We noted that children tended to not generate responses for the last items; this may have been due to lack of time. The NuC rating scale showed no effects of gender, age or grade (p > .05 for all of the variables); with MaC, however, we found gender effects (M_{Female} = 6.22, SD_{Female} = 3.14; M_{Male} = 5.05, SD_{Male} = 2.82; t(203) = -2.796, p =.006, d = .039) in favor of females. For the other variables no differences were found.

-- insert Table 1 here--

Two separate CFAs were conducted to test the fit of a one-factor model for both NuC and MaC to the data using a maximum likelihood (ML) estimator. The score for the NuC rating scale was composed of the sum of accuracy and creativity scores. We considered these scores as one due to the high correlation between them (r =.934, p < .000). Results showed that the fit of the model for NuC (χ^2(14) = 27.270, p =.018, C-MIN/df = 1.948; CFI = .945; RMSEA = .068, CI (.028 - .106)) was not as good as for MaC (χ^2(14) = 21.580, p = .088, C-MIN/df = 1.541; CFI =

.965; RMSEA = .052, CI (.000-.092)). To investigate the possibility of improving the fit of the NuC rating scale, an Exploratory Factor Analysis (EFA) was carried out using principal axis factoring with varimax rotation. Table 2 displays the results of the EFAs for both the NuC and the MaC rating scales. The EFA for NuC (KMO = .779, Bartlett's test, $\chi^2(21) = 259.775$, p = .000) revealed two factors considering the eigenvalue criteria—the first explaining 31.205%, the second 6.425%, together accounting for 37.630% of the variance. The scree plot shows the two factors. All of the factor loadings on the NuC rating scale were over .4. The factor scores were saved via the regression method and their correlation was .239 ($p < .01$). The grouping of CN items 1 and 2 in a different factor suggests that these items evaluate a distinct dimension of numerical creativity. Considering the dual nature of CN, we speculated that the first factor may be more related to numerical creative ability, and the second factor with the social creativity of the response. This will be explored further in later analyses and the discussion.

--insert Table 2 here--

The EFA for the MaC scores (KMO = .807, Bartlett's test, $\chi^2(21) = 211.4$, $p = .000$) suggested that one factor explains 30.5% of the variance, considering the eigenvalue criteria. The scree plot shows clearly one factor, and all of the factor loadings are over .4 except for item 7. The different explained variances for the NuC and MaC rating scales indicate that probably more than the one dimension explains student performance in CN.

In the next step we considered the relationship between CN and the components of the Aurora-a test. Considering the model of the Aurora Battery it was expected that the total score for both rating scales would correlate with the creative ability and number domain of Aurora-*a*'s tests. Yet, considering the low explained variance found in the EFA, it appeared possible that the factor scores and the total score might be related to other abilities or to the domain of the complete test. To assess this aspect, different bivariate correlations were calculated. The total scores of the subtests of Aurora-*a* (for both domains and abilities) were calculated by summing the scores of their respective subtests, and these scores were correlated with the factor scores for the NuC and MaC rating scales. The total scores for Creative ability and Numeric domain were computed without the NuC and MaC scores.

--insert Table 3 here--

Table 3 shows the correlations between the factors of both rating scales, and the domains and abilities of Aurora-*a*. The values of the correlations vary across abilities and domains for both rating scales. The differences between these values were checked using Fisher's *z*-test. The highest correlations for the MaC rating scale were with practical abilities and the verbal domain, while for Factor 1 (labeled 'social creativity') of the NuC rating scale, the highest correlations were in creative abilities and the word and image domains. Factor 2 (labeled 'numerical creativity') of the NuC rating scale showed a significant correlation with word domain only. Fisher's *z* test showed that all of the correlations between the MaC and NuC rating scales for Factors 1and 2 were statistically different ($p < .03$ for all the pair comparisons), except between the MaC and Factor 1 in the numerical domain and analytical and creative abilities. That is, the MaC rating scale was more strongly associated with Aurora's subtests than the NuC rating scale, especially with respect to Factor 2.

Which scoring provides more information about children's performance on academic tests?

Descriptive data for KS2 and MidYIS performance

To determine whether the performance on CN was related to the students' achievement on the KS2 and MidYIS tests, bivariate correlations and multiple regressions were calculated for each subtest and total score. The subsample that had also taken the MidYIS test was composed of 85 children, 40 males and 45 females with an age range of 9.42 to 10.33 years old (*M* = 9.8 years, *SD* = .25 years). The Skills subtest showed gender differences, with females performing better than males (M_{female} = 112.68, SD_{female} = 10.39; M_{male} = 102.02, SD_{male} = 21.19, *t*(82) = 2.987, *p* = .004, *d* = .655). For the other MidYIS subtests, no gender or age differences were found. Furthermore, in the MidYIS group, the females performed better than the males in the MaC rating scale, (M_{female} = 1.01, SD_{female} = .46; M_{male} = .73, SD_{male} = .43; *t*(83) = -2.850, *p* = .006, *d* = .625). The group who had also taken the KS2 was composed of 20 children, 9 males and 11 females. The age range was 12.08 to 12.8 years (*M* = 12.45 years, *SD* = .24 years). No gender or age differences were found. In Table 3, we report the descriptive information for each subtest and the total scores for the MidYIS and KS2 tests.

The results showed that the MaC rating scale presented significant correlations with all of the subtests and the total scores for both the MidYIS and KS2, except for the Math and Science subtests of the KS2.

The relationship of Cartoon Numbers to school achievement

As gender differences were found in the MidYIS, the regression analysis controlled for this variable only for this test. Two independent structural equation models were carried out for the NuC and MaC rating scales. The dependent variables were the subtests of the MidYIS and gender was included in the model as an independent variable for the subtests; all the error terms for the MidYIS subtests were correlated.

The model for the MaC rating scale showed an acceptable fit index ($\chi^2(4) = 6.385$, *p* = .172, C-MIN/*df* =1.596 , RMSEA = .054, CI(.000-.129), CFI = .973, NFI = .941). The significant paths in this model were for the Vocabulary (*β* = .232, *p* = .036), Math (*β* = .371, *p*

<.000) and Non Verbal (β = .299, p = .006) subtests. Also the Skills subtest was significantly correlated with gender (β = -.288, p = .002). The other paths for gender and other subtests of the MidYIS were not significant. The model for the NuC rating scale considered the two factors (the creative and social factors we found in the EFA). The fit indices were acceptable considering the sample size ($\chi^2(2)$ = 1.443, p = .486, C-MIN/df =.0721 , RMSEA = .000, CI(.000 -.126), CFI = 1, NFI = .987). The significant paths with the creative factor NuC2 were the Non Verbal (β = .220, p = .044) and Skills (β = .214, p = .041) subtests; the social factor, NuC1, showed an almost significant relationship with Vocabulary (β = .208, p = .055). In addition, the path between the Skills subtest and gender was significant (β = -.307, p = .002).

The linear regression model for the MaC rating scale explained 25.9% of the variance in the KS2 Reading subtest ($F(1,18)$ = 7.655, p = .013, β = .546), and15.9% of the variance in Writing ($F(1,18)$ = 4.602, p = .046, β = .451). For Science, Math and the total score, the linear regression models were not significant. Finally the regression analyses for the subtests of KS2 and Factors 1 and 2 of the NuC rating scale were not significant.

Discussion

In the present study, we developed and compared different scorings for the Cartoon Number subtest of the Aurora-*a* battery, a subtest that aims to assess children's creative ability in the numerical domain. The first scoring approach (NuC) focused on creative ability, where the score is based on the novelty, complexity, and elaboration of the responses by taking into account the integration of social knowledge with numerical knowledge. The second scoring approach (MaC) focused on two distinct aspects, namely students' originality and fluency in their application of mathematical concepts. Our results here clearly show different subtest performance when using the NuC scoring method versus the MaC scoring method. While both take into account appropriateness, elaboration (i.e., breadth of response), and originality, the first rewards the ability of the student to incorporate aspects of number “identity” (e.g., value, shape and mathematical relationships) within a social context. The second very differently values the ability of the student to specifically apply mathematical knowledge to explain a social situation by providing mathematical analogues. The analyses presented above provide rich information on the particular behavior of both scoring approaches and their relationships to other measures.

First, the descriptive data and the subsequent factor analyses inform our understanding of the structure of the CN test and the nature of its items. According to the descriptive data, the NuC approach yielded the highest averages in the last two items, indicating that students tended to give better (more elaborate and original) answers that integrated numerical concepts into a social context for these last items. Conversely, in the MaC scoring, the highest averages were attained for the first two items, indicating responses that offered more—and more original—explanations in the form of mathematical concepts and relationships than in others. Keeping in mind that the NuC scoring approach attempts to recognize an integration of creatively applied numerical and social elements, we propose that this performance pattern may be explained by the content of the items, which were designed to present progressively more complex and/or novel social situations and thus may draw on numeric versus social elements differentially. Specifically, Items 1 and 2 present social scenarios that are more like the given example and involve only two number “characters.” This might have inclined students to use more numerical or mathematical explanations in the first two items, as it may be easier to apply such concepts and operations to more simplistic social situations. Item 3, while involving only two numbers, presents a distinct social occasion in which no emotional elements are involved (most of the other items contain an emotional element, i.e., in which one of the number characters is happy, sad, or angry); and Items 4-7 present three number characters in more complex social or situational configurations. These more intricate situations might prove more difficult to map mathematically. Hence, it is possible that in the first two items, mathematics (or other numeric concepts) is more easily applied, whereas in the rest, more socially oriented explanations are more easily applied.

This general pattern is also reflected in the exploratory factor analysis for the NuC approach, in which the first two items reflect an ability that is distinct from the latter five items. Given the discussion above, we tentatively refer to Factor 1 of the NuC rating scale as the Social Factor, and Factor 2 as the Numeric Factor. It is notable that the confirmatory factor analysis for the MaC rating scale yielded a good fit for a single factor model. We may tentatively refer to MaC's factor as mathematical creativity. However, we do note that in both the EFAs, the explained variances are rather low, showing that there is some amount of variability in each of the scoring methods that cannot be explained by any of the main factors of one scoring method alone. It is possible that when combined, the rating scales together might explain a much greater part of the variance. Alternatively, other cognitive processes may be involved in the resolution of CN items that we did not consider here.

Next, we calculated NuC's and MaC's relationships to other subtests of Aurora and two other performance tests administered to subsets of this sample. In terms of their relatedness to Aurora, the bivariate correlations show that the MaC scorings are generally more highly and significantly related to Aurora's abilities and domains than the NuC scorings. Notably, MaC's scores show the highest correlations with Aurora's practical and creative subtests, and with the verbal domain. This may be consistent with the fact that all of Aurora's practical assessments involve applications of different types of knowledge in common situations. MaC does the same, albeit in a creative way. In addition, Factor 1 (what we posited as the more social factor) of the NuC rating scale was more closely related to Aurora's creativity dimension and the domains of words and images; Factor 2's highest significant correlation was with Aurora's word domain. These relationships imply the dominance of the verbal dimension of CN as an open-ended test requiring a written response, as scored using the NuC rating scale. Other relationships were not significant. In sum, MaC scores turn out to be more closely related to Aurora's other subtests, in general, while both rating scales appear to be significantly related to Aurora's other creativity subtests. Yet, surprisingly, MaC is slightly more correlated with Aurora's practical subtests, perhaps because it applies mathematical concepts to familiar, everyday situations or settings, as the practical subtests tend to do. Also surprising are both rating scales' lower (MaC) or non-existent (NuC) relationships to the number domain, suggesting that these assess aspects of mathematics and numeracy in ways that other number-related subtests do not.

With respect to students' performance on the KS2, the NuC scoring appears to be unrelated to academic achievement across all subjects, while the MaC scoring shows significant relationships with Reading and Writing and the total score, but no relationship to Math and Science. This may be a reflection of what is not routinely assessed in standardized achievement tests of Math and Science, which tend to draw more on memory and rote rather than creative applications of concepts. MaC's correlation with the MidYIS mathematical component, however, was higher than those with all of the other MidYIS subtests, although all of the correlations were rather moderate but significant. This seems in keeping with the intention of the MidYIS to tap into the more general categories of developed abilities and aptitude for learning beyond academic achievement. With regard to NuC, Factor 2 was related to only the Skills subtest of the MidYIS.

However, here we must recognize some of the limitations of our study. First, the MidYIS and KS2 tests were taken approximately 4-8 months before Aurora was administered. Hence, the delta time between these measurements may partly explain the lack of association between CN and other measures of academic achievement, and the differential performance of both rating scales on the KS2 and MidYIS. In addition, the subsets of children in this study who took the MidYIS and KS2 did not overlap, so group differences may explain the differential performance of the children on these two tests. Finally, it must be acknowledged that the Aurora Battery is still under development, therefore any correlations between the battery's subtests may as yet be adjusted in future revisions.

To summarize, we might state, then, that NuC scoring produces results of a more generalized nature, combining both numerical and social elements that do not appear to be closely related to academic assessments, although one factor appears to be related to creativity. MaC scoring, on the other hand, produces results that are more focused on a single factor, and are aligned a bit more closely with academic performance. That it is more related to performance in reading and writing may not be surprising as CN does involve reading and comprehending a brief story situation and then writing a response. Its low relationship to KS2 mathematics is surprising, yet reinforces the assumption that creative applications of mathematics are probably neither taught nor assessed for routinely.

Conclusion

Two types of creativity may be assessed using Aurora's Cartoon Number subtest—a more general creativity with respect to number and social concepts (NuC), and a more specific creativity regarding the application of mathematical knowledge (MaC). Although we hold both to be relevant to a better understanding of a child's range of abilities, the MaC score shows potentially important implications for the future study of creativity in mathematics. On the surface, its mechanisms appear to draw upon mathematical flexibility, fluency and original thinking (conventionally taught mathematics applications in unconventional problems), yet it shows fairly low to moderate relatedness to tests within numerical or academically oriented mathematical domains. A possible reason for this may be that this type of mathematical creative thinking is not assessed for in conventionally used mathematics assessment, nor is it generally accessed by other mathematics creativity tests presented in the literature thus far, which do not call for thinking broadly across domains.

Yet, on a final note, we propose that the MaC scoring of CN has strong educational implications beyond its relationship to other assessments, as it identifies a form of creative thinking that is essential for the development of creativity in mathematics: the analogizing of formal mathematics (learned in school) with a novel situation (social, in this case) that may lead to the generation of multiple possible mathematical relationships and concepts to "explain" the situation. The ability to use knowledge from one domain to explain or map another domain reflects not only a depth of understanding but also a creative ability to abstract that knowledge and apply it to novel questions or situations (Silver, 1987; Skemp, 1976). In their mathematical responses to CN items, as children draw upon what they know about mathematical relationships to represent or explain the social problem presented, they construct new causal relationships (by applying mathematical knowledge in a novel, metaphorical way) while at the same time re-constructing or reinforcing their conceptual knowledge about mathematics. This kind of thinking reflects a grasp of the abstract relationships that mathematics ultimately represents, and an ability to abstract and generalize mathematics knowledge such that it may be applied across domains. It can be related to concepts of different types of creativity, specifically mini-c as outlined by Kaufman and Beghetto (Beghetto & Kaufman, 2007; Kaufman & Beghetto, 2009), as well as to important concepts that may be applied in mathematics education, such as adaptive expertise (Baroody, 2003; Hatano, 2003; Hatano & Inagaki, 1986) and mathematical modeling (Barbosa, 2006; Glas, 2002; Kaiser & Sriraman, 2006; Reusser & Stebler, 1997).

To conclude, we present the Aurora subtest Cartoon Numbers as a promising tool for the assessment of creativity in the numerical and mathematical domains. In future studies, we hope to further refine its scoring methods, explore its relationships to other aspects of children's thinking, and define its role as an important educational assessment, particularly in the domain of mathematics.

References

Aiken, L. R. (1973). Ability and creativity in mathematics. *Review of Educational Research, 43*(4), 405-432.

Balka, D. S. (1974). Creative ability in mathematics. *Arithmetic Teacher, 21*, 633-636.

Barbosa, J. (2006). Mathematical modelling in classroom: a socio-critical and discursive perspective. *ZDM, 38*(3), 293-301. doi: 10.1007/bf02652812

Baroody, A. J. (2003). The development of adaptive expertise and flexibility: The integration of conceptual and procedural knowledge. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Barwell, R. (2004). What is numeracy? *For the Learning of Mathematics, 24*(1), 20-22.

Beghetto, R. A., & Kaufman, J. C. (2007). Toward a broader conception of creativity: A case for "mini-c" creativity. *Psychology of Aesthetics, Creativity, and the Arts, 1*(2), 73-79.

Bolden, D. S. (2010). Pre-service primary teachers' conceptions of creativity in mathematics. *Educational Studies in Mathematics, 73*(2), 143-157.

Centre for Evaluation and Monitoring. (2010). MidYIS assessments Retrieved August 2012, from http://www.midyis-project.org

Chart, H. E., Grigorenko, E. L., & Sternberg, R. J. (2008). Identification: the Aurora Battery. In J. A. Plucker & C. M. Callahan (Eds.), *Critical Issues and Practices in Gifted Education* (pp. 345-365). Waco, TX: Prufrock Press.

Cho, S., & Hwang, D. J. (2006). Math creative problem solving ability test for identification of the mathematically gifted. *Research in Mathematical Education, 10*(1), 2006.

DirectGov. (2009). National curriculum teacher assessments and key stage tests. *Retrieved from http:// www.direct.gov.uk/en/Parents/Schoolslearninganddevelopment/ExamsTestsAndTheCurriculum/ DG_10013041* Retrieved September 25, 2012

Getzels, J. W., & Jackson, P. W. (1962). *Creativity and intelligence: Explorations with gifted students*. New York: Wiley and Sons, Inc.

Glas, E. (2002). Klein's model of mathematical creativity. *Science & Education, 11*, 95-204.

Guilford, J. P. (1950). Creativity. *American Psychologist, 5*, 444-454.

Hadarmard, J. (1945). *The psychology of invention in the mathematical field*. New York: Dover Publications, Inc.

Hashimoto, Y. (1997). The methods for fostering creativity through mathematical problem solving. *International Journal of Mathematical Education, 29*(3), 86-87.

Hatano, G. (2003). Foreword. In A. J. Baroody & A. Dowker (Eds.), *The development of arithmetic concepts and skills*. Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Hatano, G., & Inagaki, K. (1986). Two courses in expertise. In H. Stevenson, H. Azuma & K. Hakuta (Eds.), *Child development and education in Japan*. New York: W. H. Freeman.

Haylock, D. (1987). A Framework for Assessing Mathematical Creativity in Schoolchildren. *Educational Studies in Mathematics, 18*(1), 59-74.

Haylock, D. (1997). Recognising mathematical creativity in schoolchildren. *International Journal on Mathematical Education, 29*(3), 68-74.

Kaiser, G., & Sriraman, B. (2006). A global survey of international perspectives on modelling in mathematics education. *ZDM, 38*(3), 302-310. doi: 10.1007/bf02652813

Kaufman, J. C., & Beghetto, R. A. (2009). Beyond big and little: The four c model of creativity *Review of General Psychology, 13*(1), 1-12.

Kim, H., Cho, S., & Ahn, D. (2003). Development of mathematical creative problem solving ability for identification of gifted in math. *Gifted Education International, 18*, 164-174.

Kornilov, S. A., Tan, M., Elliott, J. G., Sternberg, R. J., & Grigorenko, E. L. (2011). Gifted identification with Aurora: Widening the spotlight. *Journal of Psychological Assessment, 30*(1), 117-133.

Krutetskii, V. A. (1969). An experimental analysis of pupils' mathematical abilities *Soviet studies in the psychology of learning and teaching mathematics, vol. 2* (pp. 105-111). Stanford, CA: School Mathematics Study Group.

Krutetskii, V. A. (1976). *The psychology of mathematical abilities in school children*. Chicago: University of Chicago Press.

Leikin, R., & Lev, M. (2007). *Multiple solution tasks as a magnifying glass for observation of mathematical creativity.* Paper presented at the Proceeedings of the 31st Conference of the International Group for the Psychology of Mathematics Education, Seoul.

Lubart, T., & Guignard, J. (2004). The generality-specificity of creativity: A multivariate approach. In R. J. Sternberg, E. L. Grigorenko & J. L. Singer (Eds.), *Creativity: From potential to realization* (pp. 43-56). Washington, D. C.: American Psychological Association.

Lubart, T., Pacteau, C., Jacquet, A. Y., & Caroff, X. (2010). Children's creative potential: An empirical study of measurement issues. *Learning and Individual Differences, 20*(4), 388-392. doi: 10.1016/j.lindif.2010.02.006

Lubart, T., & Sternberg, R. J. (1994). Creativity. In R. J. Sternberg (Ed.), *Thinking and problem solving*. San Diego, CA: Academic Press.

Mann, E. L. (2006). Creativity: The essence of mathematics. *Journal for the Education of the Gifted, 30*(2), 236-260.

Mouchiroud, C., & Lubart, T. I. (2001). Children's original thinking: An empirical examination of alternative measures derived from divergent thinking tasks. *The Journal of Genetic Psychology, 162*(4), 382-401.

Pehkonen, E. (1995). Introduction: Use of open-ended problems. *ZDM, Zentralblatt für Didaktik der Mathematik, 27*(2), 43-57.

Pehkonen, E. (1997). The state-of-art in mathematical creativity. *International Journal of Mathematical Education, 29*(3), 63-67.

Poincaré, H. (1913). *Mathematical creation.* Paper presented at the Societé de Psychologie, Paris.

Polya, G. (1962). *Mathematical discovery: On understanding, learning and teaching problem solving*. Hoboken, NJ: John Wiley & Sons.

Reusser, K., & Stebler, R. (1997). Every word problem has a solution,ÄîThe social rationality of mathematical modeling in schools. *Learning and Instruction, 7*(4), 309-327. doi: 10.1016/s0959-4752(97)00014-5

Runco, M. A., Okuda, S. M., & Thurston, B. J. (1987). The psychometric properties of four systems for scoring divergent thinking tests. *Journal of Psychoeducational Assessment, 2*, 149-156.

Silver, E. A. (1987). Foundations of cognitive theory and research for mathematics problem-solving. In A. H. Schoenfeld (Ed.), *Cognitive science and mathematics education*. London: Lawrence Erlbaum Associates.

Skemp, R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching, 77*, 22-26.

Sriraman, B. (2004). The characteristics of mathematical creativity. *The Mathematics Educator*, *14*(1), 19-34.

Sriraman, B. (2005). Are giftedness and creativity synonyms in mathematics? *The Journal of Secondary Gifted Education, 17*(1), 20-36.

Sternberg, R. J. (1985). *Human abilities: An information processing approach*. Cambridge: Cambridge University Press.

Sternberg, R. J. (1996). What is mathematical thinking? In R. J. Sternberg & T. Ben-Zeev (Eds.), *The nature of mathematical thinking* (pp. 303-318). Mahwah, NJ: Lawrence Erlbaum.

Sternberg, R. J. (1999). The theory of successful intelligence. *Review of General Psychology*, *3*(4), 292-316.

Sternberg, R. J. (2006). The nature of creativity. *Creativity Research Journal*, *18*(1), 87-98.

Sternberg, R. J. (2010). Assessment of gifted students for identification purposes: New techniques for a new millennium. *Learning and Individual Differences*, *20*(4), 327-336.

Sternberg, R. J., Lubart, T. I., Kaufman, J. C., & Pretz, J. E. (2005). Creativity. In K. J. Holyoak & R. G. Morrison (Eds.), *Cambridge Handbook of Thinking and Reasoning* (pp. 351-371). Cambridge: Cambridge University Press.

Torrance, E. P. (1962). *Guiding creative talent*. Englewood Cliffs, NJ: Prentice Hall.

Torrance, E. P. (1974). *Torrance tests of creative thinking*. Lexington, MA: Personal Press.

Table 1

Item-level descriptives for the NuC and MaC rating scaless

	NuC					MaC				
Item	N	*M*	*SD*	Min	Max	**N**	*M*	*SD*	Min	Max
1	202	-.12	1.48	0	6	201	.52	1.53	0	3
2	203	-.09	1.40	0	6	203	1.25	2.09	0	6
3	202	-.28	1.50	0	5	201	.06	1.49	0	4
4	197	-.43	1.55	0	6	205	-.50	1.32	0	4
5	190	.01	1.83	0	6	188	.07	1.88	0	4
6	194	.71	1.61	0	6	193	-.43	1.43	0	3
7	193	.20	1.71	0	6	191	-.53	1.37	0	3
Total	205	3.07	0.81	-	-	205	.81	0.43	-	-

Notes. N = # of valid responses; Min = minimum observed score; Max = maximum observed score. The scores are presented as standardized scores based on mean and standard deviation of seven items for each rating scale.

Table 2

Exploratory factor analysis for the NuC and MaC scorings

	NuC			MaC	
	α if item deleted	Factor 1	Factor 2	α if item deleted	Factor 1
1	.746	.053	**.441**	.727	.434
2	.728	-.010	**.722**	.688	.659
3	.700	**.549**	.119	.705	.574
4	.716	**.474**	.094	.712	.540
5	.701	**.557**	.106	.707	.590
6	.709	**.587**	-.001	.697	.633
7	.703	**.782**	-.159	.736	.379

Notes. The bold numbers for the NuC rating scale reflect the main loadings on each factor.

Table 3

Bivariate correlations between NuC, MaC, the ability and domain scores for Aurora-a, and the KS2 and MidYIS tests

Test	Subtest	M (*SD*)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
MaC Scale	1. MaC	0.0 *(0.88)*	-																	
NuC Scale	2. Factor1	0.0 *(0.83)*	-.190*	-																
	3. Factor 2	0.0 (0.74)	-.107	.239*	-															
Aurora Domains	4. Words	5.70 *(6.40)*	**.363****	**.170***	**.140***	-														
	5. Image	3.51*(5.35)*	.341**	**.170***	.079	.561**	-													
	6. Number	1.10 *(4.05)*	.274**	.119	.070	.607**	.499**	-												
Aurora Abilities	7. Practical	3.13 *(5.24)*	**.390****	.128	.093	.765**	.717**	.679**	-											
	8. Analytical	0.94 (*5.67*)	.246*	.133	.083	.761**	.637**	.828**	.631**	-										
	9. Creative	6.23(*4.73*)	.369**	**.222***	.136	.748**	.760**	.498**	.564**	.563**	-									
KS2 Test	10. Reading	4.70 *(.57)*	.652*	.343	-.002	.735**	.680*	.508*	.597*	.634*	.786**	-								
	11. Writing	4.05 *(.69)*	.608*	.110	-.085	.811**	.439	.488*	.621*	.635*	.668*	.711**	-							
	12. Math	4.45 *(.69)*	-.019	.349	.133	.299	.213	.569*	.439	.451*	.211	.228	.173	-						
	13. Science	4 .60 *(.60)*	.304	.130	-.088	.484*	.342	.533*	.337	.606**	.479*	.554*	.436	.333	-					
	14. Total Score	22.45 *(2.37)*	.552*	.297	-.011	.781**	.592*	.685*	.665*	.788**	.727**	.882**	.794**	.516*	.764*	-				
MidYIS Test	15. Vocabulary	106.82 *(13.52)*	.266*	.182	.109	.719**	.473**	.514**	.595**	.583**	.575**	-	-	-	-	-	-			
	16. Math	108.16 *(13.31)*	.395**	.044	.183	.555**	.492**	.573**	.577**	.587**	.437**	-	-	-	-	-	.539**	-		
	17. Nonverbal	101.23 *(12.92)*	.321*	-.034	.209	.442**	.636**	.421**	.518**	.499**	.488**	-	-	-	-	-	.512**	.472*	-	
	18. Skills	107.74 *(17.08)*	.249*	-.019	.224*	.443**	.374**	.445**	.338*	.448**	.492**	-	-	-	-	-	.322*	.355*	.358*	-
	19. Total Score	105.67 *(11.77)*	.369*	.134	.163	.733**	.548**	.616**	.668**	.666**	.583**	-	-	-	-	-	.897**	.856**	.562**	.389**

Notes. Bold numbers indicate the highest correlations between the factors and the Aurora test. * $p < .05$, ** $p < .001$.