Cognitive Development 26 (2011) 127–141

Contents lists available at ScienceDirect

Cognitive Development

The influence of relational complexity and strategy selection on children's reasoning in the Latin Square Task

Patrick Perret*, Christine Bailleux, Bruno Dauvier

University of Provence, PsyCLE Center, France

ARTICLE INFO

Keywords:
Relational complexity theory
Reasoning
Strategy selection
Cognitive development
Latin Square Task

ABSTRACT

The present study focused on children's deductive reasoning when performing the Latin Square Task, an experimental task designed to explore the influence of relational complexity. Building on Birney, Halford, and Andrew's (2006) research, we created a version of the task that minimized nonrelational factors and introduced new categories of items. The results of two experiments conducted with school-aged children yielded an apparent dilution of complexity effects and suggest that specific inferential strategies can reduce the relational complexity that children need to process. A theoretical account is proposed emphasizing the influence of adaptive selection of strategies that mediate processing capacity constraints in reasoning development.

1. Introduction

1.1. Relational complexity and the development of reasoning

The present study explored the influence of relational complexity on children's deductive reasoning. Recent perspectives in adult cognitive psychology (Oberauer, Süb, Wilhelm, & Wittmann, 2008) suggest that the relational integration component of working memory constitutes a central mediator of human reasoning. In developmental research, this view also lies at the heart of relational complexity (RC) theory (Halford, Wilson, & Phillips, 1998), which states that growth in relational processing capacity with age increases the complexity of the mental models that children can form. In turn, these

* Corresponding author at: Centre PsyCLE, Université de Provence, 29, Avenue Robert Schuman, 13621 Aix en Provence Cedex 1, France.
E-mail address: Patrick.Perret@univ-provence.fr (P. Perret).

0885-2014/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cogdev.2010.12.003

changes in the complexity of mental models (from which inferences are derived) increase children's understanding of the world, through conceptual refinements and the development of reasoning. In line with other contemporary views of reasoning (Johnson-Laird, 1983), RC theory is thus based on the assumption that, when solving problems, the human mind constructs mental models intended to represent the structure of the relations involved in these problems. A second core assumption is that limited processing capacity in working memory limits the complexity of these models. The relational complexity metric applies both to the structural properties of a problem and to individual processing capacity. Relational complexity in RC theory is defined as the number of variables (or arguments) that must be related within the same cognitive representation. Unary relations are based on a single variable, binary relations on two variables, ternary relations on three, and so on.

Halford, Baker, McCreden, and Bain (2005) have established that quaternary relations are the most complex relations that adults can mentally represent and constitute the upper limit of human processing capacity. However, RC theory has identified two mechanisms that can help individuals sidestep this processing barrier, segmentation and chunking. Segmentation consists in breaking excessively complex tasks down into several steps, so as to reduce the relational complexity involved at each step. Chunking consists in compressing two or more variables into one. This mechanism both reduces the processing load and allows the newly compressed variables to form a single argument in a higher-order relation. RC theory regards segmentation and chunking as important components of expertise in a particular conceptual domain or a specific type of task.

According to RC theory, age-related changes in processing capacity are a crucial factor for cognitive development, but not the only one. Halford (1999) has repeatedly stressed that RC theory does not deny the role played by knowledge or experience. Processing capacity should be regarded as an enabling factor that gradually broadens the horizons of conceptual development and reasoning. Developmental changes in processing capacity are thought to occur according to a roughly predictable timetable, with children becoming able to process unary relations at a median age of one year, binary relations at two years, ternary relations at five years, and quaternary relations at 11 years (Andrews & Halford, 2002). This gradual growth in processing capacity enhances the number of variables that children can relate in their mental models, thereby allowing them to represent the structure of increasingly complex problems. As a consequence, their inferences in reasoning can rely on more accurate and adequate representations of the relational systems they are dealing with.

RC theory thus provides a clear framework for predicting children's performance on reasoning tasks. Performance should be determined mainly by the match (or mismatch) between the child's processing capacity and the task's relational complexity. Halford and Andrews (2004) revisited well-known developmental tasks (class inclusion, transitive inference, theory of mind) and showed that accurate analysis of relational complexity can help to explain both early success and late failures on variants of these tasks. More recently, Birney, Halford, and Andrews (2006) developed a new experimental task, the Latin Square Task, explicitly derived from RC theory and designed to test its predictions about the influence of relational complexity on deductive reasoning.

1.2. The Latin Square Task

The Latin Square Task (LST) is based on a matrix of 16 cells (4 × 4 structure) that can be filled with four different geometric shapes. The defining principle is that each shape should appear only once in each row or column. Participants are confronted with an incomplete matrix and asked to determine which of the shapes should be placed in a target cell. The items are designed so that the information already present in the relevant rows or columns of the array can direct a participant's inferences toward the right conclusion. In its defining principle, the task resembles Sudoku problems and, as such, constitutes a "puzzle of pure deduction" (Lee, Goodwin, & Johnson-Laird, 2008).

The LST has several important qualities with regard to RC theory requirements. First, the deductive mechanisms activated in this task are largely free from the influence of prior conceptual knowledge, pragmatic schemas or innate modules known to affect human reasoning and often difficult to disentangle in performance analyses. Second, the task also minimizes the amount of information to be held in memory and consequently maximizes the role of the processing (as opposed to storage) component of working memory in the determination of performance. Third, it relies on a single, simple rule (suitable for a broad range of ages and abilities) that can be applied to items of varying complexity. The relational complexity of LST items is manipulated by controlling the number of rows and columns that need to be simultaneously considered in order to choose the right shape for the target cell. "Binary items require integration of elements within either a single column or row . . . Ternary items require integration of information from both a row and column . . . For the quaternary items, solution is achieved by integrating elements across multiple rows and columns that are not necessarily fully constrained by a simple intersection (Birney et al., 2006, pp. 150–151).

Birney et al. (2006) studied the influence of relational complexity on the reasoning performance of university students and children aged 9–16. Participants completed a total of 18 items, with six items per RC category (binary, ternary or quaternary). Results, based on regression and Rasch analyses, confirmed the predictions of RC theory. Items of greater complexity were associated with more errors and longer response times.

However, the authors highlighted two instructive methodological difficulties. First, on the basis of empirical data, some items called for a reclassification (e.g., from quaternary to ternary), as participants could follow a valid but unanticipated reasoning pathway to find the solution. Using this alternative route of serial inferences meant that they encountered lower levels of relational complexity than the authors had envisaged when they designed the items. This phenomenon draws our attention to a crucial aspect of RC analysis, in that it applies to cognitive processes rather than to the task itself. As long as the items offer several possible paths to solution, accurately estimating the level of relational complexity participants are actually dealing with remains an uncertain enterprise.

Second, despite a clear complexity effect, the results indicated that factors other than dimensionality significantly contributed to the variations in performance on the LST. Response times were affected not just by RC but also by the number of empty cells in the matrix. Furthermore, the number of processing steps required to reach the solution significantly influenced both error rates and response times. For some items, intermediate empty cells had to be dealt with before the final inference concerning the target cell could be generated. Other items were single-step problems. Although this serial parameter was not explicitly controlled for in the generation of items in Birney et al.'s (2006) study, appropriate statistical analyses revealed that ability to undertake multistep reasoning made an important contribution to performance, above and beyond processing capacity. A recent study by Zhang, Xin, Lin, and Li (2009) confirmed that the number of processing steps required to find a solution in the LST accounts for a major proportion of variability in item difficulty.

On the basis of Birney et al.'s (2006) results, we designed a new version of the LST in order to control for these nonrelational factors and to further explore the influence of complexity on children's reasoning. Experiment 1 introduces this new version of the task and reports the performance of a sample of school-aged children.

2. Experiment 1

The aim of Experiment 1 was to study the effect of relational complexity on children's performance after controlling for the effects of the nonrelational factors identified in Birney et al.'s (2006) study. To this end, three main changes were made to the original task:

(1) The number of empty cells was kept constant in all the items;
(2) all the items required a single inferential step to identify the right shape for the target cell (no intermediate cells to deal with);
(3) in order to constrain reasoning pathways (and consequently limit individual variability in the inferential routes chosen by the participant), only the relevant rows and columns of the matrix were shown.

RC manipulations adhered to the principles defined by Birney et al. (2006), in that item complexity was a function of the number of rows and columns that had to be processed in parallel in order to perform the inferential step leading to the solution. We designed binary, ternary, and quaternary items (see Fig. 1 for examples of each category). In the ternary items, information from both a row and a column has to be integrated. We made an additional distinction between *secant* and *nonsecant* ternary

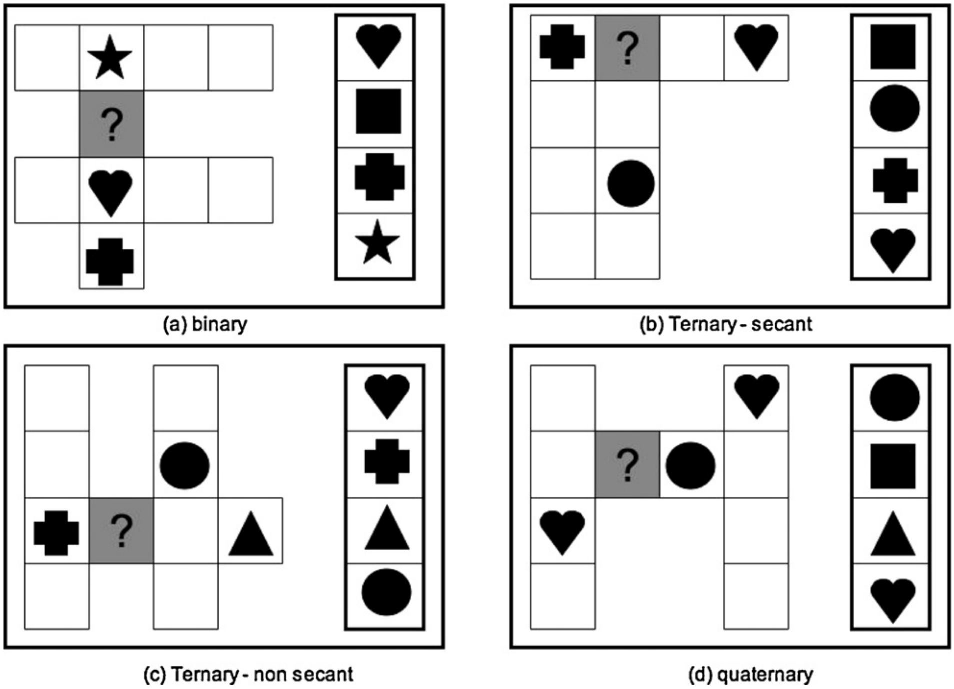

Fig. 1. Examples of the categories of items used in Experiment 1.

items, depending on the target cell's position with regard to the pieces of information that had to be integrated. As shown in Fig. 1, secant ternary items were designed so that the target cell was located at the intersection of the row and column that had to be taken into account. In nonsecant ternary items, this was not the case. A preliminary study (Perret, Bailleux, & Dauvier, 2008) had suggested that the position of the target cell in ternary items constitutes a neglected dimension of difficulty. We therefore manipulated this additional factor for ternary items, to create a total of four categories.

2.1. Method

2.1.1. Participants

Seventy-one third, fourth and fifth graders (35 girls), aged 8–11 years, took part. Mean age was 9–4 (SD = 15 months). Participants were recruited from elementary schools in a predominantly middle-class area in Aix-en-Provence, France.

2.1.2. Item generation

A set of 24 items (six binary, six secant ternary, six nonsecant ternary, and six quaternary) was generated, following the basic principles noted earlier. We used six geometric shapes (square, circle, cross, triangle, star and heart), all blue in color. The order of item presentation was randomized, with the one proviso that none of the four conditions could occur more than three times in succession. We controlled the number of times that each shape appeared in the matrix structure and in the list of response options. The position of the filled cells was also controlled. All items were based on a 4 × 4 structure, but only three rows and columns were displayed, to reduce the number of noninformative cells and possible inferential pathways. We fixed the number of filled cells at three. Consequently, the number of empty cells was kept constant. The target cell was highlighted and indicated by a question mark in the center.

2.1.3. Procedure

E-Prime software (Psychology Software Tools, Inc., 2008) was used to build our version of the LST and the test was administered by computer. Administration began with a familiarization phase featuring four sample items (binary, secant ternary, nonsecant ternary and quaternary). Participants were tested individually in a quiet room. The instruction was as follows: "*You have to find the shape missing from the cell with a question mark, obeying the following rule: Each shape must appear only once in each row and in each column. You have to choose the missing shape from these four possibilities* [participants were shown a list of four response options]. *Be careful; there is only one right answer.*" The items were displayed on the left-hand side of the computer screen and the list of response options on the right-hand side. Participants responded by clicking on one of the shapes represented in the response options. They were encouraged to do their best and to respond as accurately as possible. All participants completed the 24 items without any feedback from the experimenter.

2.2. Results

Several statistical approaches can be used to analyze successes and failures observed in a task like the LST. A simple approach is to compare the mean difficulties of groups of items in relation to their theoretical level of complexity. A gradual increase in observed difficulties as a function of relational complexity could be taken as evidence validating the theoretical analysis of the task. However, averaging across participants and items can hide differences between items and does not provide any information about the structure of individual differences. A more fine-grained approach consists in using a psychometric tool such as the Rasch model (Rasch, 1961), a type of item response theory (IRT) model. In the Rasch model, each item is defined by its own level of difficulty, making it possible to pinpoint items that deviate from the prediction. Individual differences are also taken into account, as individual abilities are assumed to be continuously distributed across a latent continuum.

We therefore adopted a twofold statistical approach. First, we used repeated-measures analysis of variance (ANOVA) to compare groups of items in a classic manner. Second, generalized linear mixed-effects models (GLMMs; Breslow & Clayton, 1993), which can be regarded as a generalization of the multilevel model (Faraway, 2005), and some IRT models such as the Rasch model (Boeck & Wilson, 2004, p. 6; Miyazaki, 2005), were fitted to the data in order to validate the item classification. In the first analysis, individual mean proportions of correct responses were computed for the binary, ternary and quaternary items. These individual values were then averaged to obtain mean accuracy per complexity level (solid line in Fig. 2). Results clearly followed the expected stepwise increase in difficulty from binary to quaternary items. A repeated-measures ANOVA revealed a relatively strong and significant effect, $F(2, 136) = 176$, $p < .001$, $\eta^2 = .55$.

Item-by-item examination of the data revealed that the group of ternary items encompassed very different levels of difficulty. It was immediately obvious that the distinction between secant and nonsecant items within the group of ternary items would have to be taken into account, as shown in Fig. 2. Hence, we distinguished between four types of item: binary (a), secant ternary (b), nonsecant ternary (c), and quaternary (d).

In order to empirically find the best form of item classification, three GLMMs (binomial distribution and logit link function) were fitted to the data, using the lme4 package (Bates & Sarkar, 2009) in R (R Development Core Team, 2009). These models are very similar to IRT models such as the Rasch model, where item difficulty is taken into account by the fixed-effects parameters and participants' abilities by the random-effects parameters (Boeck & Wilson, 2004; Doran, Bates, Bliese, & Dowling, 2007). Like IRT models, the GLMMs were directly fitted to the binary success/failure data without any averaging. This methodology offered some useful features in this context, allowing us to form groups of items and to force the difficulty parameters to the same level within a given group of items. In each of the three models, a different classification of items was used. A comparison of the models told us which classification was the most relevant, given the data.

In the first model (M1), items were classified according to their theoretical level of complexity. In the second model (M2), we added the distinction between secant ternary (b) and nonsecant ternary (c) items. In the third model (M3), binary (a) and secant ternary (b) items were set to the same level of difficulty, as were nonsecant ternary (c) and quaternary (d) items. A model comparison based

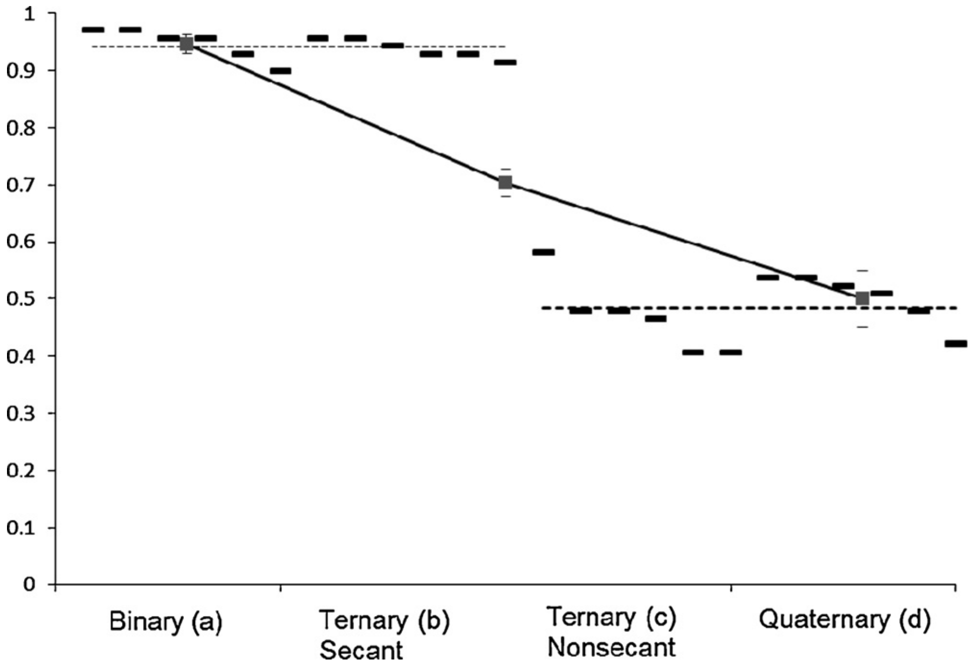

Fig. 2. Mean accuracy per item and mean accuracy for each theoretical complexity level in Experiment 1. *Note*: The proportions of correct responses were first computed for each individual participant. The solid line shows group mean accuracy with standard error for each theoretical level of complexity (binary, ternary, quaternary). Within each category of items (binary, secant ternary, nonsecant ternary, quaternary), items are ordered by proportion of correct responses. The thin dotted line represents M3 predictions.

Table 1
Goodness of fit of the generalized linear mixed effect models in Experiment 1.

Model	Item classification	DF	AIC	BIC	Subject variance
Generalized linear mixed effect models					
M1	(a) (b–c) (d)	4	1595	1617	1.50
M2	(a) (b) (c) (d)	5	1297	1324	2.41
M3	(a–b) (c–d)	3	1295	1311	2.40

Note: (a) Binary, (b) secant ternary, (c) nonsecant ternary, (d) quaternary, (b–c) all ternary items were constrained to the same level of difficulty in the model, (a–b) binary and secant ternary were constrained to the same level of difficulty in the model. The Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criiterion (BIC; Schwarz, 1978) are widely used cues for model selection. They simultaneously take the fit and the parsimony of the model into account. The lower the AIC and the BIC, the better the model.

on AIC and BIC criteria[1] (Table 1) showed that M3 was the best model (AIC = 1295, BIC = 1311). M1, which did not distinguish between ternary items, showed the poorest fit (AIC = 1595, BIC = 1617) and was unsuitable for revealing individual differences (low between-participant variance of 1.5). As M2 (AIC = 1297, BIC = 1324) was no better than M3, there seemed no empirical reason to make a distinction between binary and secant ternary or between nonsecant ternary and quaternary. The mean observed proportion of correct responses for nonsecant ternary and quaternary items (.48) was higher than it would have been if children had used a guessing strategy (.25, given four response options).

[1] The Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criiterion (BIC; Schwarz, 1978) are widely used cues for model selection. They simultaneously take the fit and the parsimony of the model into account. The lower are the AIC and the BIC, the better is the model.

2.3. Discussion

The items in our version of the Latin Square Task were created so as to minimize the impact of three factors known to have a confounding effect: the number of empty cells, the number of processing steps, and individual variability in inferential pathways. Results indicated that, when only the relational dimension of the items was considered, the data clearly confirmed the predictions of RC theory, as the mean proportion of correct responses decreased as a function of relational complexity. However, when the secant/nonsecant distinction between ternary items was taken into account, as suggested by the GLMMs, a more complex picture emerged. Binary and secant ternary items were found to be of equivalent difficulty for children, and a comparable equivalence emerged for nonsecant ternary and quaternary items. The latter results seem to argue against the moderating role of RC on performance.

A possible explanation is that the strategic dimension of the LST was neglected in previous analyses of the task. On the basis of their work on Sudoku problems, Lee et al. (2008) suggested that individuals may recruit various strategies to cope with Latin square problems and variations in performances may be rooted in strategic shifts. With regard to the present task, we hypothesized that two types of strategies may underlie children's reasoning, cell-based reasoning and shape-based reasoning. The distinction between them lies mainly in the initial focus of the child's attention – a given cell versus a given shape. Cell-based reasoning consists of (i) focusing attention on a given row or column in the array that will constitute the starting point for subsequent inferences, (ii) concentrating on a particular *cell* in this row or column, (iii) determining the shapes it could be filled with, given those already present in the same row or column, and then, if more than one possibility is found, (iv) eliminating potential shapes by taking additional constraints from intersecting rows or columns into account, until only one shape is left. In the examples of Fig. 1, the secant ternary item can be solved by (i) focusing attention on the first row of the matrix, (ii) concentrating on the target cell with a question mark, (iii) determining the shapes it could be filled with, given those already present in the row (either a circle or a square), and (iv) eliminating the circle by virtue of its occurrence in the intersecting column, so that the square is the only possibility left.

Shape-based reasoning, in contrast, consists of (i) focusing attention on a given row or column in the array that will constitute the starting point for subsequent inferences, (ii) concentrating on a particular *shape* among the response options, (iii) determining the cells where it could be placed, given those already filled in the row or column in question, and then, if there is more than one possibility, (iv) eliminating potential cells by taking into account additional constraints from intersecting rows or columns, until there is only one cell left. The quaternary item in Fig. 1, for example, can be solved by (i) focusing attention on the one row in the matrix, (ii) concentrating on the heart option, (iii) identifying the three cells where the heart could be placed, given the presence of the circle, and (iv) eliminating the two end cells by virtue of the hearts that are already present in the intersecting columns, so that the only possibility for the heart in the row in question is the cell with the question mark.

The essence of the distinction between cell-based and shape-based reasoning is the search for possible shapes for a given cell versus the search for possible cells for a given shape. We contend that this strategic distinction may help to explain the data from Experiment 1. Whereas binary and secant ternary items can be solved using cell-based strategies, nonsecant ternary and quaternary items require shape-based strategies, as can be seen in the quaternary item in Fig. 1. Once the child has identified three possible candidates to fill the target cell (the heart, the star, and the square), no additional constraint is available to disambiguate the decision. Thus, these categories of items require a shape-based strategy. Furthermore, the shape-based strategy may allow children to chunk the information available from multiple rows and columns when producing the final inference: the shape under consideration can go *nowhere else* except in the target cell. When the inference is made, the number of intersecting rows and columns that have previously been considered does not constrain the reasoning because this information has been reduced to a single argument ("The shape can go nowhere else except. . ."). This chunking mechanism could explain why complexity effects were not clearly observed in Experiment 1. As Halford and Andrews (2004) point out, ". . . it has been a major difficulty for cognitive complexity analyses that humans have very proficient strategies to reduce processing loads" (p. 129). In Experiment 2, we tested the hypothesis that performance on the LST may rely on deductive strategies that reduce the processing demands of complex items.

3. Experiment 2

Experiment 2 was designed to examine whether school-aged children can solve quinary items. We have suggested that the dilution of complexity effects in Experiment 1 may have resulted from a reduction in the processing demands of excessively complex items. If so, increasing item complexity beyond the previous four dimensions would not preclude success, even though this level of complexity clearly surpasses children's processing capacity, according to RC theory.

A further aim of Experiment 2 was to replicate the results of Experiment 1, using a visual display of the task that more closely resembled that used by Birney et al. (2006). In ad hoc interviews, some participants in Experiment 1 indicated that, as they were performing the task, they developed a heuristic that consisted of looking through the response options and choosing the shape that was missing from the array. Roberts (2000) emphasized the fact that many deduction tasks give rise to the development of such shortcut perceptual strategies in order to "obtain a solution directly from the problem statements" (p. 25). Due to our simplified display of the task, a systematic use of this heuristic (choosing the shape that is not already in the array) would have yielded accurate responses for binary and secant ternary items, although not for nonsecant ternary and quaternary ones (where the correct responses corresponded to shapes already present in the array). In Experiment 2, we sought to rule out the possibility that our results were the byproduct of a simplified display. Thus, we used a complete matrix (with five rows and columns) and increased the number of filled cells so as to avoid use of such shortcut strategies. Nevertheless, drawing on Birney et al.'s (2006) results, we retained the objective of limiting the impact of extraneous factors that were likely to have a confounding effect. To this end, the LST items in Experiment 2 had the following features:

(1) The size of the matrix was increased (5 × 5 structure) to allow the creation of quinary items (i.e., requiring consideration of the possible elements in the target row or column whilst also taking elements in four other rows or columns into consideration);
(2) the number of cells provided per item was increased, but held constant across items;
(3) in contrast to Experiment 1, a complete matrix of rows and columns was presented, but in order to limit individual variability in the inferential pathways chosen by participants, we highlighted the relevant row or column that had to be considered first;
(4) all items required a single inferential step to identify the right shape for the target cell (i.e., no intermediate cell to deal with);
(5) relational complexity was manipulated according to the principles defined by Birney et al. (2006), and four categories of items were used (with six items per category): secant ternary, nonsecant ternary, quaternary and quinary (see Fig. 3 for an example of each category). Binary items were excluded from this experiment, as a clear floor effect emerged in Experiment 1.

3.1. Method

3.1.1. Participants

Participants were 89 third, fourth and fifth graders (49 girls), aged 8–11 years, from the same population as drawn from in Experiment 1., France. Mean age was 10–1 (*SD* = 12 months).

3.1.2. Item generation

A set of 24 items (six secant ternary, six nonsecant ternary, six quaternary, and six quinary) was generated in the same way as in Experiment. In this version, we used six blue shapes (square, circle, cross, triangle, heart, and moon) and added an additional control: in the ternary items, the response was always one of the shapes already present in the array (in order to preclude the use of the heuristic described in discussion of Experiment 1). All items were based on a 5 × 5 structure. All the cells were visible, but we highlighted the most informative row or column in a pale yellow color, in order to guide participants' reasoning and reduce variability in inferential pathways. We fixed the number of filled cells at five, with the result that the number of empty cells was also constant. The target cell was indicated by a question mark and highlighted in a luminous yellow.

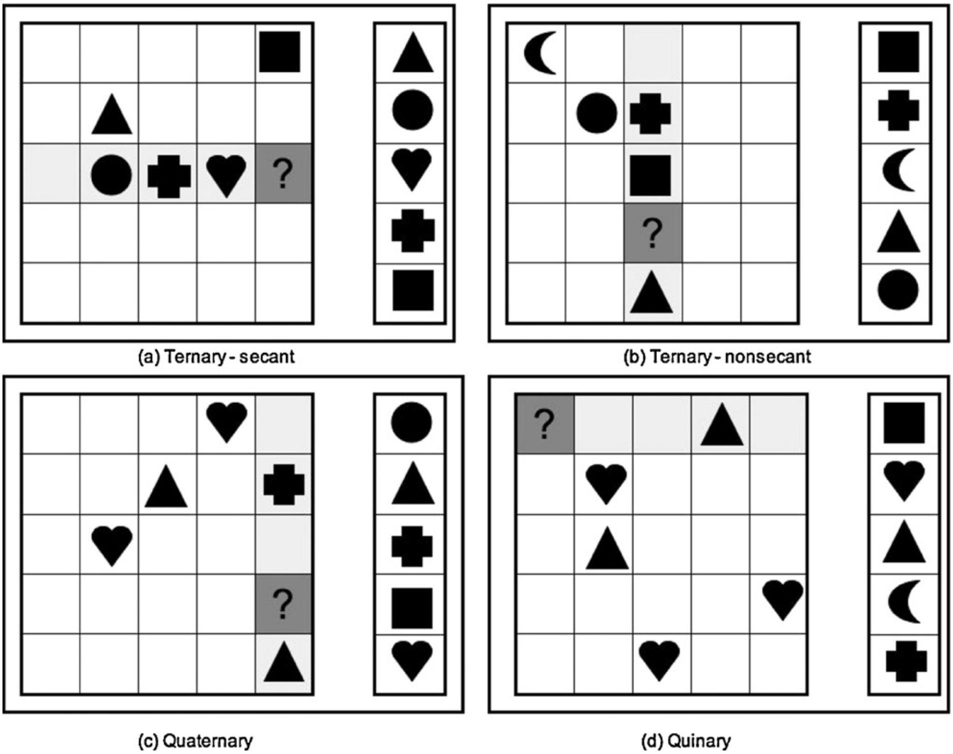

Fig. 3. Examples of the categories of items used in Experiment 2.

3.1.3. Procedure

The procedure was the same as in Experiment 1, with a familiarization phase featuring six practice items (two binary, one secant ternary, one nonsecant ternary, one quaternary and one quinary). All participants completed the 24 items, without any feedback from the experimenter.

3.2. Results

At the group level, no difference in level of difficulty was observed between the nonsecant ternary, quaternary and quinary items. Mean accuracies were very similar, at around .5. A repeated-measures ANOVA revealed a significant effect of complexity level, $F(3, 264) = 35.8$, $p < .001$, $\eta^2 = .17$, due to the high success rate observed for the secant ternary items (.84).

To ensure similar levels of difficulty within each group of items and to validate item classification, three GLMMs were fitted to the data. As shown in Fig. 4, secant ternary items appeared to have a homogeneous level of difficulty. Nonsecant ternary, quaternary and quinary items, however, showed greater within-group variability across items. In the first model (M1), items were categorized in the four categories – secant ternary, nonsecant ternary, quaternary, and quinary. In the second model (M2), nonsecant ternary, quaternary, and quinary items were grouped into a single class and their difficulty levels were constrained to equality. This classification was strongly suggested by the descriptive statistics. The third model (M3) was a Rasch model, in which each item had its own difficulty parameter. The Rasch model allowed us to test within-group variability across items. If this variability was too high, it would not be appropriate to subject the difficulty levels to a single-value constraint within a given group of items, and a more flexible model like the Rasch model would show a better fit.

The best model according to both AIC and BIC criteria was M2 (Table 2), which meant that no distinction in mean difficulty levels needed to be made between nonsecant ternary, quaternary and

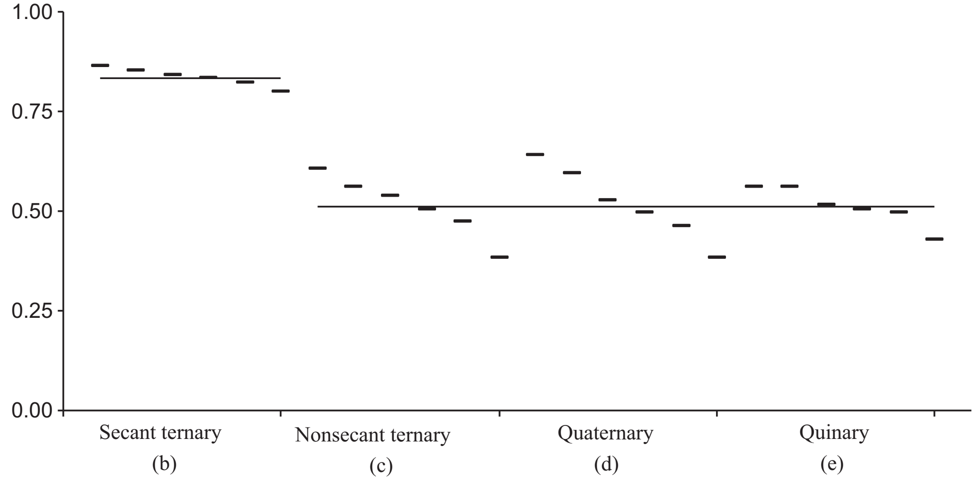

Fig. 4. Mean accuracy per item in Experiment 2. *Note*: within each complexity level, the items are shown in order of decreasing accuracy. The thin dotted line represents M2 predictions.

quinary items at the group level. The Rasch model (M3; AIC = 2439, BIC = 2581) was no better than M2 (AIC = 2438, BIC = 2455). The apparent heterogeneity shown in Fig. 4 within the nonsecant ternary, quaternary and quinary groups of items was not great enough to require a more fine-grained classification. This apparent heterogeneity was probably due to the fact that the difficulty levels of these items were distributed around .5. Variability is necessarily greater at the center of the scale than it is at the two ends.

In order to test the hypothesis regarding individual strategic differences, two additional analyses were conducted, regression analyses and finite mixtures of generalized linear regression models. Whereas the secant ternary items could be solved using a cell-based reasoning strategy, we hypothesized that the nonsecant ternary items would require a shape-based strategy comparable to the one required by the quaternary and quinary items. As a consequence, individual performances on nonsecant ternary items would be a better predictor of the performances on quaternary and quinary ones, which hypothetically involve the same inferential processes. We conduced two regression analyses, with secant and nonsecant ternary scores as predictors. The score on nonsecant ternary items significantly predicted performances on quaternary and quinary items. Standardized regression coefficients were .33 ($p = .001$) and .25 ($p = .015$), respectively. Conversely, the score on secant ternary items was not predictive, as standardized regression coefficients were not significant – .11 and .10 ($p > .05$ for both).

Table 2
Goodness of fit of the generalized linear mixed effect models and finite mixture of generalized linear models in Experiment 2.

Model	Item classification	DF	AIC	BIC	Subject Variance
Generalized linear mixed effect model					
M1	(b) (c) (d) (e)	5	2442	2470	1.24
M2	(b) (c–d–e)	3	2438	2455	1.24
M3 (Rasch)	Item by item (24)	25	2439	2581	1.31
Finite mixture of generalized linear models					
M4	(b) (c) (d) (e)	9	2373	2424	2
M5	(b) (c) (d) (e)	14	2335	2414	3
M6	(b) (c) (d) (e)	19	2298	2406	4
M7	(b) (c) (d) (e)	24	2290	2426	5

Note: (b) Secant ternary, (c) nonsecant ternary, (d) quaternary, (e) quinary, (c–d–e) nonsecant ternary, quaternary and quinary items were constrained to the same level of difficulty in the model.

Table 3
Description of the four groups revealed by latent class analyses in Experiment 2.

	Class 1	Class 2	Class 3	Class 4
Number of children	24	7	30	28
Age: mean (in month)	123.96	118.14	119.47	120.57
Age: standard deviation	9.80	5.40	13.04	12.87
Accuracy on secant ternary	.86	.24	.87	.93
Accuracy on nonsecant ternary	.77	.26	.46	.41
Accuracy on quaternary	.92	.26	.12	.67
Accuracy on quinary	.91	.31	.13	.63

The GLMMs previously fitted to the data were not necessarily appropriate for examining individual differences in strategy use. These models assume that individual differences are ordered along a latent continuum. However, if the availability of the shape-based strategy differs from one child to another, individual differences have to be viewed as qualitative differences between groups of children. We tested the hypothesis of discrete ability levels, using a type of latent class model called “finite mixtures of generalized linear regression models” (FM-GLM; Grün & Leisch, 2008). This methodology provided a means of comparing models with different numbers of classes of children and with specific levels of item difficulty within each class. We implemented the FM-GLMs using the FlexMix R package (Grün & Leisch, 2007; Leisch, 2004). Another difference between FM-GLMs and GLMMs is that in the former, the levels of item difficulty are computed within each class. Hence, there is no more absolute item characteristic. This means that two items can be ordered differently in two different classes regarding their level of difficulty. This is a useful feature, given our theoretical assumption that the perceived difficulty of, say, a quaternary item might be very different for a child using the shape-based strategy as opposed to a child whose reasoning was based solely on the cell-based strategy. Although the number of parameters increases rapidly with the number of classes, the goodness-of-fit criteria used to select the model (i.e., AIC and BIC) take parsimony into account.

Four models were fitted to the data, with between two and five classes. The four FM-GLMs all had better fits than M2, suggesting that a discrete representation of individual differences was more appropriate. According to the AIC criterion, the five-class model (M7; AIC = 2290) was slightly better than the four-class one (M6; AIC = 2298), whilst the latter exhibited the smallest BIC score (2406). The four-class model was thus selected as the best model, favoring parsimony and based therefore on the BIC criterion. Table 3 provides a description of the four groups.

The first class (C1) contained 24 children, who achieved high accuracy levels on every item, even the quinary ones (accuracy ranging from .77 to .91). A smaller class of seven children (C2) comprised participants who exhibited a low level of accuracy on all four item types. The accuracy levels we observed in this second class, ranging from .24 to .31, were close to the expected chance score of .25. The third class (C3) contained 30 children who correctly solved secant ternary items most of the time (accuracy of .87) but had far greater difficulty with nonsecant ternary items (.46) and with quaternary and quinary ones (.12 and .13). The last class (C4) was composed of 28 children who performed very well on secant ternary items (.93), but moderately so on nonsecant ternary items (.41), like the children in C3. However, the children in C4 differed from their C3 counterparts on quaternary and quinary items, as they solved them most of the time (accuracies of .67 and .63).

C1 showed the highest mean age (Table 3), followed by C4, then C3, and finally C2, corresponding to children who performed at chance level. However, age differences across classes were not significant, $F(3, 85) = .8$, $p < 1$, *ns*.

3.3. *Discussion*

Although we modified the visual display of the task so that it precluded recourse to a shortcut heuristic, the results of Experiment 2 broadly replicated those of Experiment 1. A significant difference in children’s performances once again emerged between the two categories of ternary items (secant versus nonsecant). Furthermore, the model of item classification that best fitted the data made no distinction between nonsecant ternary, quaternary, and quinary items with regard to their levels of difficulty. This result, in conjunction with the relatively high success rate for quinary items, confirmed the hypothesis that specific deductive mechanisms can free performance from the constraint of limited processing capacity. A central assumption of RC theory is that quaternary relations constitute the upper limit of human processing capacity. Beyond this level of complexity, segmentation or chunking strategies are needed to reduce the task's processing load. The fact that performance on nonsecant ternary items significantly predicted performance on quaternary and quinary ones further suggests that children's reasoning on those items relied on comparable mechanisms of complexity reduction.

The chunking mechanisms involved in the categories of items represented in Fig. 3 can be expressed with reference to Birney et al.'s (2006) formalization principles. Chunks are indicated by continuous underlining, and cells in the array are coded as follows (the first digit indicates the row number and the second digit the column number).

Secant ternary: not ($\underline{\text{circle, cross, heart}}$, square) → triangle (cell-based reasoning);
Nonsecant ternary: not ($\underline{\text{cross, square, triangle}}$, moon in 1.3) → moon in 4.3 (shape-based reasoning);
Quaternary: not ($\underline{\text{cross, triangle, heart in 1.5 or 3.5}}$) → heart in 4.5 (shape-based reasoning);
Quinary: not (triangle, $\underline{\text{heart in 1.2, 1.3 or 1.5}}$) → heart in 1.1 (shape-based reasoning).

It can be concluded from these RC analyses that nonsecant ternary, quaternary, and quinary items can be chunked to equivalent levels of complexity (i.e., ternary), despite differences in dimensionality; further, they differ from secant ternary items by virtue of a qualitatively different inferential process (shape-based reasoning).

Another key aspect of the present results is the latent class analyses. These revealed four different patterns of performance, clearly suggesting individual variability in the way children approached the four categories of items. In line with the hypothesis formulated in Experiment 1, these performance profiles may reflect the differential availability of cell- and shape-based strategies to support inferences, with shape-based strategies representing a more sophisticated form of reasoning. A possible explanation is that shape-based reasoning involves an additional inferential step: if a shape cannot occur in any of the other empty cells in a row or column, it has to occur in the cell in question. Within this framework, the four groups show different levels of strategy mastery. Class 2 children were unable to apply either of the deductive strategies to this task. Class 3 children were able to use cell-based reasoning for secant ternary items, but failed to apply or switch to a shape-based strategy when confronted with nonsecant ternary, quaternary, or quinary items. Class 1 children were able to switch efficiently between these two strategies according to item requirements.

Children in Class 4 presented the most surprising pattern of performance, with a U-shaped curve for performance as a function of relational complexity. They succeeded better on quaternary and quinary items than on nonsecant ternary ones. One interpretation of this pattern is that the quaternary and quinary items evoked a shape-based strategy more strongly, insofar as the shape that children needed to take into account appeared two or three times in the array and may therefore have attracted their attention. A rather similar phenomenon was reported by Lee et al. (2008) in their work on Sudoku problems. They noticed that, despite an overall complexity effect, an increase in relational complexity was not always associated with an increase in item difficulty. "The departure from the effects of relational complexity when its value was five is probably attributable to the ease of noticing four instances of the same digit in columns and rows intersecting the box containing the target cell" (p. 353).

Interestingly, the four groups of children identified by the latent class analyses did not significantly differ in age. Given the age range of our sample, developmental changes in processing capacity between 8 and 12 years should have induced age-related latent classes if relational complexity was the only factor influencing performance. Conversely, if specific inferential strategies allowed a reduction in the relational complexity that children needed to process, a developmental trend would be less likely. Our results suggest that there are no clear developmental constraints to preclude the emergence of either cell- or shape-based strategies, at least for children in the age range we examined.

4. General discussion

Drawing on Birney et al.'s (2006) research, the present study further documented the influence of relational complexity on children's deductive reasoning in the Latin Square Task. In relation to previous versions of the task (Birney et al., 2006; Zhang et al., 2009), two main changes were introduced. First, we controlled for the effects of nonrelational factors that were likely to have confounding effects. Second, we introduced a distinction between two categories of ternary items, based on the target cell's position with regard to the information that needed to be integrated – secant versus nonsecant ternary items. Compared with previous research, the results of both experiments revealed an apparent dilution of complexity effects: items of the same theoretical level of complexity were characterized by different levels of difficulty, and items of different theoretical levels of complexity were characterized by comparable levels of difficulty. The discrepancy appears because RC analyses are concerned with cognitive processes rather than items' features. Investigation of required inferential processes suggested that a classification scheme based on items' dimensionality does not necessarily match the complexity of the relations actually processed. Children's mean performances on the so-called "quinary" items in Experiment 2 (which were comparable to their performances on nonsecant ternary and quaternary items) confirmed that their inferences were based on reasoning mechanisms that reduced the task's theoretical processing load.

Furthermore, latent class analyses revealed the existence of considerable individual variability in response patterns for the various categories of items. Thus, any theoretical account of the present data needs to explain not only why one category of items was more difficult than another, but also why one particular pattern applied to some children but not to others. We propose that children's reasoning on the LST relied on two deductive strategies, focusing either on the possible shapes for a given cell or on the possible cells for a given shape. We argue that the latter strategy reduced the items' theoretical processing load because the final inferential step allowed previous information to be chunked into a single argument. This chunking mechanism could explain the dilution of complexity effects observed in both our experiments. This account also provides a possible explanation for individual variability in response patterns, as reflecting the differential availability and/or flexibility of use of the cell- and shape-based reasoning strategies.

Two implications, one methodological the other theoretical, can be drawn from the present study. In a methodological vein, Birney et al. (2006) regarded the LST as a promising psychometric tool for assessing the influence of the processing component of working memory. We share the view that future explorations of how working memory influences the development of reasoning would benefit from a task specifically designed to assess processing capacity. However, given the present data, it may be premature to recommend the use of the LST in this regard, for unless chunking strategies are controlled for, performance on the task cannot be regarded as a valid measure of processing capacity. This challenge was clearly expressed by Halford et al. (2005), as quoted earlier. Halford and colleagues have stressed that the relevance of RC analyses depends on the accuracy of independently constructed process models of the tasks. In this regard, the present study may contribute to our understanding of the cognitive processes used by children in the LST and provide useful constraints for future analyses of complexity.

At a more theoretical level, Markovits and Barrouillet (2004) have called for a revival in the developmental study of reasoning and put forward three possible theoretical frameworks within which to address the question of "what develops" in children's reasoning: (a) accounts that emphasize age-related changes in executive control, (b) metalogical approaches that regard children's progress as the expression of an enhanced understanding of their own inferences, and (c) neo-Piagetian perspectives, according to which age-related increases in attentional resources constitute a key factor. In the light of the present discussion, a fourth approach clearly needs to be mentioned: the adaptive selection of strategies from the child's repertoire (Kuhn & Pearsall, 1998; Siegler, 1996). Indeed, we interpreted individual profiles as the expression of the differential availability and/or flexibility of use of the cell- and shape-based strategies.

Psychologists studying reasoning have often regarded variations in strategy use as a noisy phenomenon obscuring the identification of universal inferential processes (Roberts, 2000). By contrast, Siegler (2007) assumes adaptive selection of strategies as at the core of cognitive growth: "Recognizing this variability is important not only for accurately describing development but also for understanding cognitive change" (p. 104). Studies of strategy development need to focus on the mechanisms responsible for both the emergence of new strategies and the inhibition of older ones (Kuhn & Pease, 2010). The strategy selection process can be viewed as governed by two major influences (Reuchlin, 1978) – internal and external constraints. External constraints, or affordances (Gibson, 1977), correspond to problem features (whether relevant cues or misleading superficial properties) that make the use of a particular strategy more likely at the group level. Internal constraints reflect the fact that for a given individual, some strategies are more easily activated than others. These two kinds of constraints interact during the strategy selection process and can give rise to different kind of mistakes. Microgenetic studies (Flynn & Siegler, 2007; Kuhn & Pease, 2010) should explore both the validity of the strategic distinction suggested here and the extent to which strategic flexibility contributes to changes in performance on the LST.

Although the aforementioned four different approaches to reasoning development have given rise to specific research programs in the past, they are in no way mutually exclusive in their accounts of the changes that affect children's reasoning. RC theory, for example, does not preclude strategic or knowledge effects but works in combination with them (Halford & Andrews, 2004). Given the clear interaction between several factors influencing children's performances on the LST, we contend that this task could constitute a fruitful experimental context for studying their dynamic interactions. The present study highlights how the strategic dimension of the task influences the relational complexity that children need to process and, as such, mediates processing load constraints. Future research could focus on how far children's adaptive shifting from one strategy to another as a function of the items' requirements (i.e., the executive dimension of the task) relies on an increased awareness of their own inferential processes, that is, on metalogical development.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Andrews, G., & Halford, G. S. (2002). A cognitive complexity metric applied to cognitive development. *Cognitive Psychology*, *45*, 153–219.

Bates, D., & Sarkar, D. (2009). *lme4: Linear mixed-effects models using S4 classes*. Retrieved from: http://CRAN.R-project.org/.

Birney, D. P., Halford, G. S., & Andrews, G. (2006). Measuring the influence of complexity on relational reasoning: The development of the Latin Square Task. *Educational and Psychological Measurement*, *66*, 146–171.

Boeck, P. D., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9–25.

Doran, H., Bates, D., Bliese, P., & Dowling, M. (2007). Estimating the multilevel Rasch model: With the lme4 package. *Journal of Statistical Software*, *20*(2)

Faraway, J. J. (2005). *Extending the linear model with R: Generalized linear, mixed effects and nonparametric regression models*. New York: Chapman & Hall/CRC.

Flynn, E., & Siegler, R. (2007). Measuring change: Current trends and future directions in microgenetic research. *Infant and Child Development*, *16*, 135–149.

Gibson, J. J. (1977). The theory of affordances. In R. E. Shaw, & J. Bransford (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 67–82). Hillsdale, NJ: Lawrence Erlbaum Associates.

Grün, B., & Leisch, F. (2008). Finite mixtures of generalized linear regression models. In C. Shalabh, & Heumann (Eds.), *Recent advances in linear models and related areas* (pp. 205–230). Berlin: Springer.

Grün, B., & Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis*, *51*(11), 5247–5252.

Halford, G. S. (1999). The development of intelligence includes the capacity to process relations of greater complexity. In M. Anderson (Ed.), *The development of intelligence* (pp. 193–213). Hove, UK: Psychology Press.

Halford, G. S., & Andrews, G. (2004). The development of deductive reasoning: How important is complexity? *Thinking and Reasoning*, *10*, 123–145.

Halford, G. S., Baker, R., McCredden, J. E., & Bain, J. D. (2005). How many variables can humans process? *Psychological Science*, *16*, 70–76.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences*, *21*, 803–865.

Johnson-Laird, P. N. (1983). *Mental models*. Cambridge, UK: Cambridge University Press.

Kuhn, D., & Pearsall, S. (1998). Relations between metastrategic knowledge and strategic performance. *Cognitive Development*, *13*, 227–247.

Kuhn, D., & Pease, M. (2010). The dual components of developing strategy use: Production and inhibition. In H. Waters, & W. Schneider (Eds.), *Metacognition, strategy use, and instruction* (pp. 135–159). New York: Guilford Press.

Lee, N. Y., Goodwin, G. P., & Johnson-Laird, P. N. (2008). The psychological puzzle of Sudoku. *Thinking & Reasoning*, *14*, 342–364.

Leisch, F. (2004). {FlexMix}: A general framework for finite mixture models and latent class regression in *R. Journal of Statistical Software*, *11*(8), 1–18.

Markovits, H., & Barrouillet, P. (2004). Introduction: Why is understanding the development of reasoning important? *Thinking and Reasoning*, *10*, 113–121.

Miyazaki, Y. (2005). Some links between classical and modern test theory via the two-level hierarchical generalized linear model. *Journal of Applied Measurement*, *6*(3), 289–310.

Oberauer, K., Süb, H.-M., Wilhelm, O., & Wittmann, W. W. (2008). Which working memory functions predict intelligence? *Intelligence*, *36*, 641–652.

Perret, P., Bailleux, C., & Dauvier, B. (2008). The Latin Square Task and the measure of processing capacity in RC theory. In *Paper presented at the 18th advanced course of the archives Jean Piaget on cognitive development: Mechanisms and constraints* Geneva, Switzerland.

R Development Core Team (2009). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from: http://www.R-project.org.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability* (pp. 321–334). Berkeley: University of Chicago Press.

Reuchlin, M. (1978). Processus vicariants et différences individuelles. *Journal de Psychologie Normale et Pathologique*, *75*(2), 133–145.

Roberts, M. J. (2000). Individual differences in reasoning strategies: A problem to solve or an opportunity to seize? In W. Schaeken, G. De Vooght, A. Vandierendonck, & G. d'Ydewalle (Eds.), *Deductive reasoning and strategies* (pp. 23–48). Mahwah, NJ: Lawrence Erlbaum.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464.

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.

Siegler, R. S. (2007). Cognitive variability. *Developmental Science*, *10*, 104–109.

Zhang, L., Xin, Z., Lin, C., & Li, H. (2009). The complexity of the Latin Square Task and its influence on children's performance. *Chinese Science Bulletin*, *54*, 766–775.