Measuring the Influence of Complexity on Relational Reasoning

The Development of the Latin Square Task

Damian P. Birney
*University of Sydney, Australia*

Graeme S. Halford
*University of Queensland, Brisbane, Australia*

Glenda Andrews
*Griffith University, Brisbane, Australia*

Relational complexity (RC) theory conceptualizes an individual’s processing capacity and a task’s complexity along a common ordinal metric. The authors describe the development of the Latin Square Task (LST) that assesses the influence of RC on reasoning. The LST minimizes the role of knowledge and storage capacity and thus refines the identification of a processing-capacity-related complexity effect in task performance. The LST is novel with one explicit rule that is easily understood by adults and children. In two studies, a test of 18 items encompassing three RC levels was administered to university (*N* = 73; 16-33 years) and school (*N* = 204; 8-19 years) students. Rasch analyses indicate that the LST scores were psychometrically stable across age groups and provides important diagnostic clues for task development. Consistent with RC theory, the LST is sensitive to parallel and serial (via segmentation) processing demands. The LST provides a strong basis for research on working memory and related constructs (fluid intelligence).

**Keywords:** working-memory; relational complexity; reasoning; Rasch analysis; Latin square

The assessment of cognitive complexity and its impact on performance has consistently been of theoretical and applied interest in education and psychology. Relational complexity (RC) theory was proposed by Halford (1993) and Halford, Wilson, and Phillips (1998a) as a solution to the problem of how to define complexity in cognitive tasks independent of domain. RC theory conceptualizes processing capacity of the individual and task complexity along a common ordinal metric. A number of advances have been made toward the quantification of cognitive complexity using the RC approach (e.g., Andrews & Halford, 1998, 2002; Birney & Halford, 2002; Halford, Andrews, Dalton, Boag, & Zielinski, 2002; Halford, Baker,
McCredden, & Bain, 2005; Halford et al., 1998a). In the current article, we report on the development of the Latin Square Task (LST), a new experimental task that assesses processing components of working memory and is derived explicitly from the theory of RC. The Method for Analysis of Relational Complexity (MARC) is applied to derive items with predicted a priori psychometric properties, and these predictions are extensively tested using Rasch and classical test theory (CTT) methodologies in two populations—university students and school-aged children. First, we consider the theoretical foundations of the RC theory.

Relational Complexity

A common assumption in cognitive psychology is that increasing task demands results in a concomitant increase in the resource required for successful performance. However, there are many ways to degrade performance without necessarily increasing the “depth” of processing that is needed. For instance, one can increase the storage demands of an otherwise simple task and make success unlikely because it is impossible to keep all the necessary information active long enough to make a decision. Miller (1956) introduced the concept of chunking to explain how people can overcome the apparent limitation in storage capacity. However, storage does not seem to be the only or even the most important piece of the cognitive puzzle. Halford et al. (1998a) argued that it is the complexity of the relations between the pieces of information that is subject to processing capacity limitations, and not the amount of information per se.

The ability to deal with cognitive complexity has also been associated with intelligence (Marshalek, Lohman, & Snow, 1983), and this link has been proposed as a method to distinguish between manipulations of complexity and manipulations of difficulty generated by other factors (Spilsbury, Stankov, & Roberts, 1990; Stankov, 2000). That is, correlations with measures of fluid intelligence should increase with task complexity all else being equal, not with increases in difficulty. This criterion is purely empirical in that it is agnostic to the underlying reasoning processes that might be entailed. RC theory provides a complementary approach to describing person and task characteristics that influence the complexity of reasoning.

Specification of RC Theory

Relational reasoning and its link with RC theory can be derived from a small number of premises. First, reasoning entails representing and processing relations between task entities. Second, processing internally represented relations generates nontrivial cognitive demand. Third, the complexity of the internally represented relation can be used to quantify the characteristics of the processes used in performing the task—this is RC. And fourth, processing capacity is a function of the peak RC of the representations that an individual can process, and therefore, processing capacity can be quantified using the same metric. This reasoning can be summarized by two axioms that form the basis of the RC theory. Axiom 1: Complexity of a cognitive process “is the
number of interacting variables that must be represented in parallel to implement that process” (Halford et al., 1998a, p. 805).

**Axiom 2:** Processing complexity of a task “is the number of interacting variables that must be represented in parallel to perform the most complex process involved in the task, using the least demanding strategy available to humans for that task” (Halford et al., 1998a, p. 805).

In general, the complexity of a relation, \( R(a_1, a_2, \ldots, a_n) \), is determined by the number of arguments, \( n \). Each argument in a relation \( (a_1, a_2, \ldots, a_n) \) is a source of variation in that it can be instantiated in more than one way under the condition that the relation is true (or at least perceived to be true). Unary relations have a single argument as in class membership, dog(Fido). Binary relations have two arguments as in larger-than(elephant, mouse). Ternary relations have three arguments as in arithmetic-addition(2,3,5), and so on. The more arguments that are required to instantiate a relation, the greater the cognitive complexity of the task and the greater the demand placed on working memory.

Given that processing capacity limitations constrain the amount of information that can be represented in parallel, an individual needs to be able to work within these limits. RC theory proposes that cognitive demand can be reduced through the processes of conceptual chunking and segmentation. Conceptual chunking is the recoding of a relation into a lower dimensional concept. For instance, velocity can be considered as a function of distance and time (velocity = distance/time) and in this form entails a ternary relation, ratio(distance, time, velocity). However, velocity can also be considered a unary relation, such as velocity(60km/hr), as in the position of a speedometer needle. Conceptual chunking reduces processing demand, but at the cost that relations between chunked variables become inaccessible to the current reasoning process. While we are thinking of speed as a unary relation, questions about time and distance cannot be considered.

Segmentation entails decomposing problems with many arguments into a series of lower dimensional processes that are solved in series; this is the link between serial processing and RC theory. Relations are only defined between variables that are in the same segment (i.e., step) and relations between variables in different segments are inaccessible to the current reasoning process. It follows that variables can be chunked or segmented for a given step in a task only if relations between them do not need to be processed for that step. The complexity of the most complex segment determines the RC of the overall task. Defined in this way, RC captures the peak cognitive requirements of the task as a whole (Halford et al., 1998a). Together, these basic principles form the foundations of the MARC.

The RC approach/theory has been successfully applied in recent years to topics in cognitive development (Andrews & Halford, 1998, 2002; Andrews, Halford, Bunch, Bowden, & Jones, 2003; Halford et al., 2002), logical reasoning in adults (Birney & Halford, 2002; Halford et al., 2005), and applied areas such as mathematics education (English & Halford, 1995) and air traffic control (Boag, Härtel, & Halford, in press). Evidence for RC theory has also come from dual-task studies (Foley & Berch, 1997; Maybery, Bain, & Halford, 1986) and neurological research that has associated relational reasoning with increased activation in regions of the brain associated with ex-
executive function, the main processing component of working memory (Waltz et al., 1999).

One of the major disagreements of the interpretation of the evidence presented in support of RC theory centers on the role of knowledge in processing and the determination of task complexity. For instance, Sweller (1998) raised concerns about the influence of knowledge on performance and of individual differences in chunking and segmentation. That is, to what extent do different segmentation strategies (either taught or acquired) change the effective RC of a task, and how can differential use of strategies be identified?

In the current article, we endeavor to further clarify the application of MARC while addressing some of these concerns. We test the theory by using RC principles to explicitly develop a constrained reasoning task with a priori specifications of cognitive demand. We wanted a task that minimizes confounds that might weaken the identification of a complexity effect. Factors such as storage load and prior experience are addressed by developing a task that is novel to participants, which (a) has a small number of general rules that can be mastered quickly by people of differing ages and abilities, and (b) has no explicit rules that are specific to a complexity level. In the remainder of this article, we outline the development of the LST. We then summarize two studies that were designed to explore the internal consistency of our RC analyses and the psychometric properties of the measures obtained from the LST in two populations: university students and school children.

Development of the Latin Square Task

The Latin square derives its name from an ancient puzzle that dealt with the number of ways Latin letters could be arranged in a square table so that each letter appeared only once in every row or column (this is the defining principle). The Latin square was first explicated for experimental use in agriculture by R. A. Fisher (1925) to control statistically for soil variability and was “discovered” by psychology in the late 1930s to facilitate interpretations of analysis of variance (e.g., Thomson, 1941). In the LST, partial (incomplete) \(4 \times 4\) Latin squares are presented and reasoners determine which of four possible elements should fill a target cell so that the matrix satisfies the defining principle.

Manipulating Cognitive Complexity in the LST

It can be demonstrated (Birney, 2002, p. 73) that when given three appropriate elements in a single row or column of a \(4 \times 4\) matrix, the status of the unknown fourth element can be determined uniquely. In fact, the value of a cell can be determined without necessarily knowing or considering all elements in a given dimension (i.e., row or column). Some combination of knowledge of row and column elements can be used in certain cases to derive the value of the target cell. For example, consider Figure 1 (Panel 1). On the left, a complete Latin square is presented, and on the right, a partial Latin square in which the values of the nondiagonal cells are hidden. We are interested
in determining the value of the target cell, D4. From just the information in the partial square, the values of each of the hidden cells cannot be determined uniquely, yet we know unambiguously that the target cell must be a 1 (we will return to a similar example in more detail shortly). There are at least three main ways that the Latin square principle can be instantiated, and we will show that this serves as the basis for the complexity manipulation.

**Binary processing.** Binary items require integration of elements within either a single column or row but not across both. In the example problem in Figure 1 (Binary), three elements in Column C are given. The fourth can be determined by comparing the elements that are present in Column C with the four elements known to complete the
set. There is no need to consider any other cells in the square. Using simple conjunction (AND) and implication (→) relations and the notation outlined by Birney and Halford (2002), this can be represented as $\text{AND}(C1(4), C3(2), C4(3)) \rightarrow C2(1)$, which has an RC = 2 and is read “C1 is a 4 AND C3 is a 2 AND C4 is a 3 implies C2 is a 1”. Underlining is used to represent the separate chunked arguments.

The logic underlying the binary classification might be thought of in a slightly different way. In Figure 1 (Binary), the elements that should exist in every row and column are represented by the numerals 1 through 4. This information is provided in the task instructions and response options. In this case, we are provided with the elements {2}, {3}, and {4} in Rows 3, 4, and 1 of Column C, respectively. This implies two sets of elements—the complete known set {1,2,3,4} and the given set {2,3,4}. Consistent with the principle of conceptual chunking discussed earlier, the comparison of these two groups of elements entails a binary relation. The relations between elements within the chunks do not need to be considered to make the current decision. We need only know that element {1} is different from all {2,3,4}, and we are not concerned with relations between {2}, {3}, and {4}. Therefore, according to the principles of MARC, they can be chunked. This will become clearer in the ternary and quaternary problems when constraints on chunking are imposed.

**Ternary processing.** Ternary items require integration of information from both a row and column. In the example of Figure 1 (Ternary), the solution is achieved through the integration of information from Column B and Row 2. The intersection, B2, must not contain an element that is present, or can be determined to be present, in the other cells of Column B or Row 2, and therefore the target response is element {3}. The representation of this ternary process can be written as $\text{AND}(B1(1), B4(4), D2(2)) \rightarrow B2(3)$, which has an RC = 3. The arguments of the conjunction cannot be combined into a single argument in this case because they require integration across more than one dimension. That is, the {2} in D2 cannot be chunked with the other terms of the conjunction. The reason is that by the Latin square principle, elements in Row 2 are not independent of the elements in any of the intersecting columns. The cell that intersects with Column B is the target cell, so elements in Row 2 need to be considered to make the current decision. The relation between the known elements in Column B does not need to be considered per se, and therefore the elements can be chunked in accordance with the principles of MARC.

**Quaternary processing.** 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. In the example problem in Figure 1 (Quaternary), the target cell cannot be determined by the binary and ternary strategies just described. These less complex strategies result in knowing only that the target cell is not a {3}. We can determine that A4 cannot be a {1} because there is already a {1} present in Column A (i.e., in A1). Similarly, C4 cannot be a {1} because Column C already contains this element (i.e., in C3). The value of D4 is determined because it contains a {3}.
By definition, a \{1\} must be present in Row 4, and therefore B4 must be a \{1\}. In this example, we are considering all possible elements in Row 4, while taking into consideration the elements in Columns A, B, and C. We are also taking into consideration the distribution of a particular element, \{1\}. The requirement to consider multiple rows and columns is the general principle that quaternary items are based on. Here it is very clear that relations between the elements need to be considered and cannot be chunked as in the binary and ternary examples.

This protocol can be thought of as the integration of three pieces of information: First, A1(1) \rightarrow \text{NOT}(A4(1)); second, C3(1) \rightarrow \text{NOT}(C4(1)); and third, D4(3) \rightarrow \text{NOT}(D4(1)). An integration of these three pieces of information is necessary to reach the implication that B4 must be a \{1\}, and this can be represented as follows:

\[
\text{AND}(\text{A1(1) \rightarrow NOT(A4(1))}, \text{C3(1) \rightarrow NOT(C4(1))}, \text{D4(3) \rightarrow NOT(D4(1))}) \rightarrow \text{B4(1)},
\]

A simplified representation is: \text{AND}(\text{A1(1), C3(1), D4(3)}) \rightarrow \text{B4(1)}, which has an RC = 4.

Note that an alternative analysis based on determining elements in Column B rather than Row 4 results in the same classification through a different but isomorphic line of logic. That is, A1(1) precludes the possibility of B1(1). Similarly, B2(3) necessarily precludes the possibility of B2(1), and C3(1) precludes the possibility of B3(1). The integration of this information still implies that B4 must be a \{1\}, and although this strategy is different to the first, it involves the same level of processing.

In summary, the complexity of the processing required in the LST depends on how the given elements need to be integrated to arrive at a solution. Items can also be constructed such that the elements of empty nontarget cells must be resolved before the target cell can be determined. This introduces a manipulation of serial processing because each of the parallel-processing steps needs to be resolved to facilitate determination of the subsequent steps.

Empirical Test of the Complexity Manipulation

To empirically test the RC analysis, a series of 18 items was generated and administered to university students. This is the first empirical presentation of this task, and we spend considerable time exploring the data from this sample. We are particularly interested in the differential difficulty of items as a function of the a priori complexity classification. Our analyses therefore focus on the psychometric properties of items (difficulty and response times). As a replication, the same 18-item test was administered to primary and high school students. The focus of the analyses for this sample is to consider the generalizability of the results obtained from the university sample with a much more academically diverse population and with a much wider age range.
Method

Participants

The first sample was composed of 73 university students enrolled in a 1st-year undergraduate subject at the University of Queensland, Australia. The students received 1% course credit for their participation. The second sample of participants comprised 210 school students recruited from two schools (one primary and one secondary) in the same suburb of Brisbane, Australia. The students classified as Year 9 and 10 in Table 1 were actually Year 8 and 9 students (respectively) tested in the last week of the academic year. All other students were tested in the 1st week of the following academic year (i.e., 8 weeks later). A small number of students did not complete all the items in the LST, and therefore the analyses are based on 204 subjects. The school students were transported to the university for testing and were administered a variety of other cognitive tasks (“puzzles”) as part of a larger testing program. To help maintain appropriate levels of motivation, school students and accompanying teachers were provided lunch and other physical activities (i.e., games and tour of the university lakes). Only the results of the LST are reported here. Table 1 summarizes the demographic characteristics of the two samples.

Table 1
Demographic Characteristics of the University and School Samples

<table>
<thead>
<tr>
<th>Academic Year</th>
<th>Classification</th>
<th>Mean Age (SD)</th>
<th>Male</th>
<th>Female</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>University</td>
<td>First-year psychology</td>
<td>18.62 (2.87)</td>
<td>13</td>
<td>60</td>
<td>73</td>
</tr>
<tr>
<td>School</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Years 4 and 5</td>
<td>Lower primary</td>
<td>9.16 (0.58)</td>
<td>26</td>
<td>29</td>
<td>55</td>
</tr>
<tr>
<td>Years 6 and 7</td>
<td>Upper primary</td>
<td>11.27 (0.67)</td>
<td>22</td>
<td>36</td>
<td>58</td>
</tr>
<tr>
<td>Years 8, 9, and 10</td>
<td>Lower secondary</td>
<td>13.57 (0.80)</td>
<td>27</td>
<td>37</td>
<td>64</td>
</tr>
<tr>
<td>Years 11 and 12</td>
<td>Upper secondary</td>
<td>16.24 (1.02)</td>
<td>8</td>
<td>19</td>
<td>27</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>12.08 (2.46)</td>
<td>83</td>
<td>121</td>
<td>204</td>
</tr>
</tbody>
</table>

Note: University age range = 16-33 years. School age range = 8-19 years.

Item Generation

A set of 6 binary, 6 ternary, and 6 quaternary items was generated following the basic principles outlined above. All items followed the $4 \times 4$ structure, and the four unique elements were drawn randomly for each item from six possible shapes (circle, square, triangle, cross, diamond, and no-shape) and six possible colors (red, green, blue, cyan, magenta, and no-color). The no-color and no-shape options were not paired for obvious reasons. Item complexity was determined by the complexity of the most complex process within an item. No attempt was made to equate (or experimentally manipulate) the number or complexity of additional processing steps within an
item. Table 2 lists the overall complexity classification and the number of incomplete cells (INC) for each item. It also provides a summary of the complexity of any additional processing steps (2D, 3D, 4D). For instance, consider Item 10 in Table 2. This item has been classified as ternary, and 6 of the 16 cells are incomplete. There are two binary processes (2D) and one ternary process (3D), and the item is therefore a three-step ternary problem. Actual items are available from the first author on request.

Procedure

Participants were tested in groups of 1 to 8. The items were presented using computers fitted with 14-inch displays. Participants were encouraged to do their best and instructed to work as quickly and as accurately as possible and to do all their working in their heads. A series of four practice items oriented participants to the nature of the task and how to make a response. The incomplete Latin square item was always presented towards the left hand side of the display screen, and the complete list of the four unique response options was provided on the right (see Figure 1 for a sample formatted item). Column and row labels were only provided during practice to facilitate detailed feedback for both incorrect and correct responses. Participants indicated which element should fill the marked cell by clicking on one of the possible responses. Following practice, the 18 test items were presented in a different random order to each student without feedback.

Results and Discussion for University Sample

As the Latin Square Task was developed using the principles of RC theory, the critical test of the complexity analysis is the performance of each item. For instance, if an item classified as quaternary turns out to be solved correctly by a large number of participants, this would suggest that we have overlooked a less complex solution strategy in our original classifications. The item set would need to be refined and tested further. We begin with item analyses employing both traditional and Rasch methodologies. We then consider the RC classification as a function of item difficulty and response time.

Item Analyses: A Rasch Approach

The 18-item test was submitted to a Rasch analysis as implemented in the RUMM software (Andrich, Lyne, Sheridan, & Luo, 1998). To take full advantage of the scaling properties of item response theory (IRT), a close fit of the data to the Rasch model needs to be observed (Hambleton & Swaminathan, 1985). When fit to the model is good, item difficulty and person ability are calibrated on the same interval scale (Andrich, 1988; Wright, 1999). Preliminary checks on the data resulted in Item 4 being excluded from the calibration analysis because it was answered correctly by all participants and therefore the item provides no information about individual ability. In subsequent analyses, we adopt the approach advocated by Green and Smith (1987)
Table 2
Latin Square Item Characteristics: Complexity, Number of Processing Steps, and Incomplete Cells

| Item | INC | 2D | 3D | 4D | Σ  | Item | INC | 2D | 3D | 4D | Σ  | Item | INC | 2D | 3D | 4D | Σ  |
|------|-----|----|----|----|----|------|-----|----|----|----|----|----|------|-----|----|----|----|----|----|
| 1.   | 9   | 1  | 0  | 0  | 1  | 7.   | 8   | 0  | 1  | 0  | 1  | 13.  | 11  | 0  | 0  | 1  | 1  | 1   |
| 2.   | 8   | 1  | 0  | 0  | 1  | 8.   | 9   | 0  | 1  | 0  | 1  | 14.  | 9   | 0  | 1  | 1  | 2  | 2   |
| 3.   | 9   | 2  | 0  | 0  | 2  | 9.   | 9   | 1  | 1  | 0  | 2  | 15.  | 11  | 0  | 0  | 1  | 1  | 3   |
| 4.   | 8   | 2  | 0  | 0  | 2  | 10.  | 7   | 2  | 1  | 0  | 3  | 16.a | 10  | 0  | 0  | 1  | 1  | 4   |
| 5.   | 7   | 3  | 0  | 0  | 3  | 11.  | 6   | 1  | 2  | 0  | 3  | 17.  | 10  | 0  | 0  | 1  | 1  | 5   |
| 6.   | 8   | 3  | 0  | 0  | 3  | 12.  | 10  | 0  | 1  | 0  | 1  | 18.b | 9   | 0  | 0  | 1  | 1  | 6   |

Note: INC = number of incomplete cells (including the target); Σ = total number of processing steps.

a. Reclassification of relational complexity (RC) Item 16: INC = 10; 2D = 1; 3D = 1; 4D = 0; Σ = 2.
b. Reclassification of RC Item 18: INC = 9; 2D = 1; 3D = 2; 4D = 0; Σ = 3.
and assign Item 4 a value of one standard error of measurement below the easiest validly calibrated item (i.e., Item 2). A further restriction of the Rasch analysis is that calibration of person ability is not possible for individuals with extreme scores. As a result, 2 participants who answered every item correctly were omitted. Table 3 lists the item difficulty estimates, a selection of fit statistics based on the Rasch model for the university sample, and the corresponding proportion correct and response time statistics and point-biserial item-total correlations. We consider item-fit and person-fit, two commonly accepted ways of assessing fit. The tests of fit are based on the 17 items and 71 university subjects.

Item fit. The most common indicator of item fit is the infit statistic—the mean squared standardized residual for each item across all individuals, weighted by specific information about the individual and the item (i.e., response variation) (see Masters & Wright, 1996). Low infit values suggest that the data fits the model too well—that there is a deficiency in the stochastic variation that is necessary for useful measurement (i.e., responses are too consistent). High infit values suggest there is more noise in the data than what has been modeled (Linacre & Wright, 1994). Although there is no hard rule for determining when fit is too low or too high, some proponents of the Rasch methodology suggest that for multichoice tests in which the “stakes” for accurate response are normal (i.e., not high), a reasonable range is between 0.70 and 1.30 (e.g., Wright, Linacre, Gustafson, & Martin-Lof, 1994). The infit statistics shown in Table 3 are all within this range, suggesting appropriate fit at the item level. The Rasch item-separation reliability is .91.

Person fit. An aggregated global measure of person fit is person separation, which is somewhat analogous to the classical reliability coefficient. A person separation of .56 obtained in the university sample indicates that the test as a whole provides somewhat modest discrimination between participants. Person-infit values can also be derived (Masters & Wright, 1996) and help place the aggregated person separation index in context. An analysis of person fit indicated that 87.3% of the sample was within the acceptable infit region. Only 4.2% of the sample had infit values greater than 1.30; the remaining students falling outside the fit region had low values, suggesting their response was too consistent with the Rasch model.

Given the modest person-separation (reliability), it is instructive to consider more closely the implications of low and high person-fit statistics. An example of an actual response pattern (ordered by calibrated item difficulty from easiest to hardest) where person-infit is under the lower limit (infit = 0.54) is 11111111111100000. This student has been flagged as misfitting because her response was “too” consistent, yet psychologically this seems a reasonable response—the student performed well on the easier binary and ternary items but failed most of the quaternary items. An example where person-infit was substantially above the upper limit (infit = 1.77) was the response pattern, 10000000011100110. This student failed most of the easier items but managed to get two of the most difficult and a few of the moderately difficult items correct. We would conclude that this student has not been measured well.
### Table 3
Descriptive Statistics for Rasch and Classical Analyses for University and School Samples

<table>
<thead>
<tr>
<th></th>
<th>University Sample (N = 71)</th>
<th></th>
<th></th>
<th></th>
<th>School Sample (N = 204)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Difficulty</td>
<td>SE</td>
<td>Infit</td>
<td>p</td>
<td>Mean RT (SD)</td>
<td>rpb</td>
<td>Difficulty</td>
<td>SE</td>
</tr>
<tr>
<td>1</td>
<td>-2.029</td>
<td>0.56</td>
<td>0.98</td>
<td>0.94</td>
<td>11.97 (26.11)</td>
<td>0.34</td>
<td>-1.158</td>
<td>0.18</td>
</tr>
<tr>
<td>2</td>
<td>-3.455</td>
<td>1.06</td>
<td>1.08</td>
<td>0.99</td>
<td>7.80 (3.61)</td>
<td>0.14</td>
<td>-1.327</td>
<td>0.19</td>
</tr>
<tr>
<td>3</td>
<td>-0.127</td>
<td>0.30</td>
<td>0.83</td>
<td>0.78</td>
<td>17.09 (20.55)</td>
<td>0.41</td>
<td>-0.229</td>
<td>0.16</td>
</tr>
<tr>
<td>4</td>
<td>-4.515^a</td>
<td>—</td>
<td>—</td>
<td>1.00</td>
<td>12.05 (6.75)</td>
<td>—</td>
<td>-1.635</td>
<td>0.20</td>
</tr>
<tr>
<td>5</td>
<td>-0.110</td>
<td>0.30</td>
<td>1.04</td>
<td>0.79</td>
<td>24.76 (20.95)</td>
<td>0.08</td>
<td>-0.187</td>
<td>0.16</td>
</tr>
<tr>
<td>6</td>
<td>0.216</td>
<td>0.28</td>
<td>0.87</td>
<td>0.70</td>
<td>18.80 (12.75)</td>
<td>0.42</td>
<td>0.842</td>
<td>0.16</td>
</tr>
<tr>
<td>7</td>
<td>-0.819</td>
<td>0.36</td>
<td>0.95</td>
<td>0.87</td>
<td>20.20 (21.94)</td>
<td>0.12</td>
<td>-1.044</td>
<td>0.18</td>
</tr>
<tr>
<td>8</td>
<td>-0.662</td>
<td>0.34</td>
<td>0.93</td>
<td>0.86</td>
<td>24.11 (33.84)</td>
<td>0.15</td>
<td>-0.940</td>
<td>0.17</td>
</tr>
<tr>
<td>9</td>
<td>-0.037</td>
<td>0.30</td>
<td>1.04</td>
<td>0.78</td>
<td>19.73 (11.68)</td>
<td>0.13</td>
<td>-0.333</td>
<td>0.16</td>
</tr>
<tr>
<td>10</td>
<td>0.404</td>
<td>0.28</td>
<td>1.15</td>
<td>0.68</td>
<td>33.05 (22.70)</td>
<td>0.05</td>
<td>0.101</td>
<td>0.16</td>
</tr>
<tr>
<td>11</td>
<td>0.570</td>
<td>0.27</td>
<td>1.15</td>
<td>0.65</td>
<td>49.75 (48.75)</td>
<td>0.10</td>
<td>0.336</td>
<td>0.16</td>
</tr>
<tr>
<td>12</td>
<td>-0.833</td>
<td>0.36</td>
<td>1.00</td>
<td>0.87</td>
<td>20.08 (21.37)</td>
<td>0.10</td>
<td>-1.172</td>
<td>0.18</td>
</tr>
<tr>
<td>13</td>
<td>1.225</td>
<td>0.26</td>
<td>0.93</td>
<td>0.51</td>
<td>24.27 (17.98)</td>
<td>0.37</td>
<td>0.934</td>
<td>0.16</td>
</tr>
<tr>
<td>14</td>
<td>1.767</td>
<td>0.27</td>
<td>1.15</td>
<td>0.37</td>
<td>56.92 (53.66)</td>
<td>0.13</td>
<td>1.658</td>
<td>0.18</td>
</tr>
<tr>
<td>15</td>
<td>1.440</td>
<td>0.26</td>
<td>1.03</td>
<td>0.45</td>
<td>28.20 (27.77)</td>
<td>0.22</td>
<td>1.320</td>
<td>0.17</td>
</tr>
<tr>
<td>16</td>
<td>1.165</td>
<td>0.26</td>
<td>0.77</td>
<td>0.54</td>
<td>21.44 (16.91)</td>
<td>0.53</td>
<td>0.968</td>
<td>0.16</td>
</tr>
<tr>
<td>17</td>
<td>1.243</td>
<td>0.26</td>
<td>0.86</td>
<td>0.51</td>
<td>29.07 (30.36)</td>
<td>0.40</td>
<td>1.699</td>
<td>0.18</td>
</tr>
<tr>
<td>18</td>
<td>0.04</td>
<td>0.29</td>
<td>0.90</td>
<td>0.75</td>
<td>27.23 (22.54)</td>
<td>0.33</td>
<td>0.167</td>
<td>0.16</td>
</tr>
</tbody>
</table>

Note: \( p \) = proportion correct; RT = response time; \( rpb^* \) = point-biserial correlation for combined sample.

a. Difficulty was estimated—see the text for details.
nostic advantage of the Rasch model becomes clear when we realize that this information is not directly available under CTT and is certainly not available in global measures of reliability.

On the whole, the measures of fit indicate that individual performance on the LST has a satisfactory fit to the Rasch measurement model. Less than 5% of students had poor-fitting (high) response patterns, and therefore all participants were retained in the subsequent analyses. Satisfactory fit allows us some confidence in using the calibrated item information to further explore the characteristics of the LST (Wright, 1999) and to consider ways in which we can better account for the processing demands imposed by the task (Embretson, 1993, 1998).

**Item Difficulty and RC**

The relative difficulty of each item is predicted in advance by RC theory, and this is the next area of analyses to be considered. Figure 2 plots item difficulty estimates derived from the Rasch approach. By rescaling item difficulty as a function of the ability of the individual and the difficulty level of other items, measurement limitations in the proportion correct score that results in ceiling and floor effects can be somewhat overcome (Green & Smith, 1987). As an example of scaling differences in the two metrics (Rasch logits and proportion correct), and Wright’s (1999, p. 70) argument that “raw scores are not measures” considers the relative difficulty of two sets of items; Items 1 and 12, and Items 5 and 8. When difficulty is defined by proportion correct, the differences between items in each set are almost the same, about 0.07 proportional points. When defined by logits, the difference between Items 1 and 12 is 1.196 logits, whereas the difference between Items 5 and 8 is almost half of this, 0.552 logits. When the data fit the Rasch model, item calibration serves to convert the nonlinear raw scores into linear measures that have the desirable properties of conjoint additivity (see also Perline, Wright, & Wainer, 1979). Given the high accuracy on binary and ternary items, this transformation also results in better item separation.
Complexity Classification

Although a general separation between items based on the three levels of complexity is apparent in Figure 2, a closer examination of the plot of item locations reveals some interesting anomalies. The most obvious is the placement of Item 18 (classified as a quaternary item) centrally within the cluster of ternary items. A second investigation of the complexity analysis of this item indicated that although the item could certainly be solved using a quaternary process, it could also be solved using a string of two ternary processes separated by a simple binary process (i.e., three segments or steps). We follow Halford et al.’s (1998a) definition of task complexity as “the most complex process involved in the task using the least demanding strategy available to humans for that task” (p. 805). Hence, Item 18 is reclassified as a ternary item for subsequent analyses.

This misclassification raised some concerns about the remaining items, so we had the complexity analyses tested once more by a graduate student who had not been involved with the task development but was conversant in the MARC methodology. Having explained our classification scheme that is outlined above, she detected a previously unforeseen solution strategy for Item 16 that altered the classification from single-step quaternary item to a two-step ternary item. In this case, the data suggest that Item 16 is a relatively difficult item (although easier than the remaining quaternary items) and as such there was no obvious data-driven rationale for revising the analysis as there was for Item 18.

A question that arises is how do these misclassifications influence our test of the RC theory? In his review of the RC theory, Sweller (1998) stated that it is not always clear which strategy will be chosen and how segmentation might influence performance. The current misclassification might provide evidence for this view and it could be argued that application of RC theory is in fact difficult to achieve. However, it is important to note that this is not a burden on the RC theory. Halford, Wilson, and Phillips (1998b) stated that RC can account for task difficulty when applied to the processes that have been used. Halford et al. acknowledged, as did Sweller, that determining these processes can entail considerable work. We take these current misclassifications of solution strategy not as a failure of the theory nor in our application of it, but as an important part of our developing understanding of the processes and strategies entailed in this new task. However, we do acknowledge that revised analyses based on empirical data do require further empirical testing to verify that the process analyses are valid. As we will demonstrate below, there is independent empirical evidence for the revised complexity analysis in the second sample of school children.

Serial Processing

A second, more pressing issue is the overlap between binary and ternary items. This was not predicted as part of the task development. A closer inspection of the item analysis reported in Table 3 (see also Figure 2) indicates a distinct separation between Items 1 and 2 (and 4) and Items 5, 3, and 6, although all have been classified as binary. As shown in Table 2, the easier items (1 and 2) only require a single binary process.
(i.e., one step), whereas the other binary items involve two or more binary processes. A similar pattern of difficulty based on number of processing steps occurs in the ternary and quaternary problems. The psychometrically easier ternary items (8, 7, and 12) entail only one ternary process, whereas the items with three processes (10 and 11) tend to be calibrated as more difficult. Item 9, which entails two processes lies somewhere between the one- and three-step clusters of ternary items. Item 14 was the only quaternary item that entailed more than one process (i.e., two steps) and was the most difficult in the set of 18 items. The remaining single-step quaternary items tend to be grouped together at a relatively easier level of difficulty. This overall pattern of findings suggests that performance on LST problems is not simply a function of the RC of the most complex process. It seems that the ability for serial processing also influences performance on this task. The following section considers this in more detail.

Decomposing Item Difficulty: RC and Processing Steps

Although we have a reasonably consistent experimental manipulation of RC, the conditions for an analysis of processing steps are not ideal. The criterion used to generate the current Latin square items did not explicitly take into consideration the number of processes involved in the analysis. Potential ways to decompose difficulty estimates into the effects of particular cognitive components has been extensively explored in the IRT literature (e.g., Embretson, 1996; Green & Kluever, 1992; Sheehan, 1997). One approach is to regress the Rasch item difficulties on to a set of weights or frequencies that reflect one or more cognitive components (Embretson & Reise, 2000; Green & Smith, 1987). This avoids the problems with the ANOVA approach when the factors or independent variables are not independent, as is the case here. This regression approach produces estimates of the cognitive components that differed little from the sample-size intensive procedures that are based on extensions of the Rasch model (e.g., Logistic Latent Trait Model; see Green & Smith, 1987).

As reported in Table 4, RC is a very good zero-order predictor of item difficulty, such that an increase in RC tends to be associated with an increase in item difficulty, \( r(16) = .71, p < .001 \). This provides further support for the RC manipulation. The zero-order correlation between item difficulty and number of processing steps (STEPS), however, was low and not statistically significant, \( r(16) = .18, p = .49 \). The relationship between the item’s RC and STEPS was not statistically significant although the trend was negative, \( r(16) = -.30, p = .22 \). This relationship between the predictors is an artifact of the way the quaternary items have been constructed in that they mostly require one process. The less complex items tend to entail more steps.

Item difficulties. A multiple regression analysis in which the item difficulty estimates derived from the Rasch analysis were regressed on to the item’s RC (RC = 2, 3, or 4) and number of processing steps (STEPS = 1, 2, or 3), indicated that RC uniquely accounts for 64% of the variation in item difficulty, \( \beta = .84, r(15) = 5.33, p < .001 \). After controlling for the effect of RC, STEPS also becomes a significant predictor
<table>
<thead>
<tr>
<th></th>
<th>University Mean (SD)</th>
<th>School Mean (SD)</th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
<th>University Mean (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difficulty</td>
<td>0.00 (1.07)</td>
<td>—</td>
<td>0.66*</td>
<td>0.69*</td>
<td>0.71**</td>
<td>0.18</td>
<td>0.32</td>
<td>—0.25 (1.67)</td>
<td></td>
</tr>
<tr>
<td>RT</td>
<td>13.99 (3.42)</td>
<td>0.64*</td>
<td>—</td>
<td>0.99**</td>
<td>0.61*</td>
<td>0.35</td>
<td>—0.17</td>
<td>24.81 (12.25)</td>
<td></td>
</tr>
<tr>
<td>CRT</td>
<td>17.05 (6.20)</td>
<td>0.78**</td>
<td>0.94**</td>
<td>—</td>
<td>0.64*</td>
<td>0.33</td>
<td>—0.10</td>
<td>26.22 (14.04)</td>
<td></td>
</tr>
<tr>
<td>RC</td>
<td>2.89 (0.76)</td>
<td>0.68*</td>
<td>0.64*</td>
<td>0.72**</td>
<td>—</td>
<td>—0.30</td>
<td>0.55*</td>
<td>2.89 (0.76)</td>
<td></td>
</tr>
<tr>
<td>STEPS</td>
<td>1.83 (0.86)</td>
<td>0.19</td>
<td>0.38</td>
<td>0.22</td>
<td>—0.30</td>
<td>—</td>
<td>—0.64*</td>
<td>1.83 (0.86)</td>
<td></td>
</tr>
<tr>
<td>INC</td>
<td>8.78 (1.35)</td>
<td>0.34</td>
<td>—0.11</td>
<td>0.13</td>
<td>0.55*</td>
<td>—0.64*</td>
<td>—</td>
<td>8.78 (1.35)</td>
<td></td>
</tr>
</tbody>
</table>

Note: RT = response time; CRT = correct response time; STEPS = number of processing steps; INC = number of incomplete cells.
a. Item 4’s calibrated difficulty is estimated in the university sample.
*p < .05. **p < .001.
uniquely accounting for 16% of the variation in item difficulty, $\beta = .43$, $t(15) = 2.77$, $p = .016$ (this relationship was not mediated by a significant interaction, $R^2_{\text{change}} = .04$, $F[1, 14] = 1.74$, $p = .209$). The combined unique contribution of RC and STEPS (80%) exceeds the total variation accounted for by the complete model, $R^2 = .67$; adjusted $R^2 = .62$; $F(2, 15) = 14.91$, $p < .001$. Cohen and Cohen (1975) stated that this is a necessary consequence and indication of the presence of “statistical suppression” when two predictors are involved. Typically, statistical suppression is considered as the effect of one variable suppressing irrelevant variation in a second variable and as a result, this second variable becomes a significant predictor of the criterion (Cohen & Cohen, 1975; Pedhazuer & Schmelkin, 1991). In the current context, whereas the relationship between the predictors is a result of the item generation procedure, their relationship with item difficulty is not an artifact. Item difficulty is clearly a function of RC, and for the current selection of items the number of processing steps alone is not a good predictor. However, when the effect of RC is taken into consideration, the number of processing steps does have a significant influence on the difficulty of the item—such that an increase in the number of processing steps is associated with an increase in item difficulty.

**Number of incomplete cells.** It may be the case that the more elements that are displayed in the Latin square, the easier the items are regardless of their complexity. To explore this possibility, the number of incomplete cells (INC) was included as an additional predictor in the standard regression analyses (see INC in Table 2 and Table 4). The zero-order relationship between item difficulty and INC was not statistically significant, $r(16) = .318$, $p = .20$; and INC did not provide incremental prediction over and above RC and STEPS, $R^2_{\text{change}} = .04$, $F(1, 14) = 1.90$, $p = .189$. This is an interesting result because the complexity manipulation relies on resolving the ambiguity in one or more cells. The depth of this ambiguity is necessarily greater as RC increases because a larger number of elements are required for resolution. The problem we face in generating LST items is that we need to display enough elements to facilitate solution and at the same time minimize the likelihood that effort will be diverted to a time-consuming search for the value of irrelevant cells. It seems that with this set of items, the number of incomplete cells does not contribute to item difficulty beyond what RC has already achieved.

**Item Response Time and RC**

It is instructive to also consider an additional measure of cognitive demand, the time to respond to the items as a function of RC. The descriptive statistics and the zero-order correlations between the RT measures for the current items are in Table 4. RC is a statistically significant zero-order predictor of overall response time (RT) and correct response time (CRT). STEPS is not a significant zero-order predictor of any of the response time measures. These results are consistent with those observed in the Rasch based measures of item difficulty. An analysis of the correct response times did not change the nature of the interpretation of the results and therefore is not reported.
The number of processing steps becomes a statistically significant predictor of RT only once RC is taken into consideration, $\beta = .58$, $t(15) = 3.80$, $p < .001$. Together, RC and STEPS accounts for 68% of the variation in RT, $R^2 = .68$, $F(2, 15) = 15.73$, $p < .001$. This effect was not moderated by an interaction between the predictors, $R^2_{\text{change}} = .05$, $F(1, 14) = 2.35$, $p = .15$. Although the zero-order correlations between INC and RT were not statistically significant (Table 4), INC was a significant predictor of RT once RC and STEPS had been considered, $R^2_{\text{change}} = .11$, $F(1, 14) = 7.66$, $p = .02$. Hence, whereas INC does not seem to influence the difficulty of the items over and above the effect of RC and STEPS, it does seem to influence the length of time required to respond to the item. Response times tend to be longer when INC is greater.

**Results and Discussion for School Sample**

The focus of these analyses is the generalizability of the university results to a much more diverse population. The item statistics are provided in Table 3. The analysis indicates a satisfactory level of person separation (separation index = .72). The item infit statistics reported in Table 3 indicate that all items were within the acceptable range for infit, $1.00 \pm .30$ (Wright et al., 1994). The Rasch item-separation reliability = .98. The rank ordering of calibrated item difficulty scores was also highly consistent with those reported for the university sample (Spearman’s rho = .97) and this provides initial support for generalizability of the LST and the reclassification of the RC of Items 16 and 18. It is important to note that consistency in the rank ordering of the items is also significant psychometrically. It could be argued from Table 3 that there was a ceiling effect on the easier items for the university sample, and a floor effect on the difficult items for the school sample (although point-biserial correlations, $r_{pb}$, suggest the items are still useful). Maintenance of the rank order of items across these distinct populations suggests that the modest person-separation/reliability may be more a function of participant sampling than of poor item construction. Indeed, when the two samples are combined, thus increasing the range of performances, Cronbach’s $\alpha$ reliability is .76 ($r_{pb}^*$ for the combined sample are reported in Table 3 as $r_{pb}^*$). Under these conditions, the Spearman-Brown estimate of reliability for a 36-item version of the LST would be .86 (Pedhazuer & Schmelkin, 1991).

An analysis of person-fit indicated that 79.2% of the sample was within the acceptable infit region. Only 6.9% of the sample had infit values greater than 1.30 and 50% of these (i.e., 8 students) came from the lower primary school group (the youngest age group). Overall, this provides support that school-aged individuals also tend to be well measured by the LST.

**Item-Based Regression Analyses**

**Item difficulty.** To explore the RC classification, the same series of analyses conducted on the university sample was repeated for this sample. In all cases, the interpretation is identical to the university sample, and therefore we only summarize the
results briefly. The descriptive statistics and zero-order correlations are summarized in Table 4 (below the diagonal). RC was a statistically significant zero-order predictor of item difficulty. The number of processing steps was not a statistically significant zero-order predictor of item difficulty but consistent with the university sample, it did contribute to a statistically significant proportion of the variation once RC had been controlled for, $R^2_{\text{change}} = .17, F(1, 15) = 6.85, p = .02$. Together, RC and number of processing steps accounted for 62.7% of the variation in item difficulty, $F(2, 15) = 12.63, p = .001$ (adjusted $R^2 = .58$). This relationship was not mediated by an interaction between complexity and number of processing steps, $R^2_{\text{change}} = .07, F(1, 14) = 3.12, p = .10$; nor did the number of incomplete cells contribute significant unique variance over and above that accounted for by RC and number of processing steps, $R^2_{\text{change}} = .07, F(1, 14) = 3.15, p = .10$.

**Item response time.** As reported in Table 4, RC was a statistically significant zero-order predictor of mean item response time. The number of processing steps was not a statistically significant zero-order predictor but did contribute to a statistically significant proportion of the variation in mean response time once RC had been controlled for, $R^2_{\text{change}} = .36, F(1, 15) = 22.15, p < .001$. Together, RC and number of processing steps accounted for 76.0% (adjusted $R^2 = .73$) of the variation in mean item response time, $F(2, 15) = 23.70, p < .001$. This relationship was not mediated by an interaction between complexity and number of processing steps, $R^2_{\text{change}} = .04, F(1, 14) = 3.14, p = .10$. The number of incomplete cells contributed marginally to the prediction of item response time over and above that accounted for by RC and number of processing steps, $R^2_{\text{change}} = .05, F(1, 14) = 4.08, p = .06$. These results are also consistent with the findings from the university sample.

**Comparison of University and School Populations**

Our main objective in the following analyses is to consider more closely the differences between the university and school students’ performance on the LST at an item-based level. Two measures of performance were considered, item accuracy and overall item response time. The descriptive statistics for each of the 18 items for each sample are provided in Table 3. The analyses reported below follow the rationale of profile analysis using the univariate repeated-measures approach (Tabachnick & Fidell, 1989). Individual-based data is explored to avoid potential problems associated with aggregating individual differences that can arise in item-based analyses. This was deemed particularly important because we did not wish to make any a priori assumptions that the item variances of the DVs were homogenous (at least for the response time measures).

**Item Accuracy**

The univariate approach entails a mixed repeated-measures/between subjects ANOVA (18 Items $\times$ 2 Groups) to explore performance on individual items as a func-
tion of university-school grouping (adjusting for violations of the sphericity assumption in the following analyses did not change the interpretations, hence, results for the unadjusted analyses are reported). The objective was to consider the degree of consistency in the pattern of results across items and groups, and therefore the effects of interest are the main effect of group membership and the two-way interaction between groups and items. As expected, a statistically significant overall difference between the items was observed, \( F(17, 4675) = 47.44, MSE = .17, p < .001 \). We have effectively explored these item differences as a function of RC and number of steps in the item-based regression analyses reported in the preceding sections, and therefore we do not repeat them here. The main effect for group membership was also statistically significant, \( F(1, 275) = 58.88, MSE = .58, p < .001 \), such that university students tended to make fewer errors than school students. The magnitude of this effect was consistent across all items as indicated by the statistically nonsignificant interaction between items and group, \( F(17, 4,675) = 1.12, MSE = .17, p = .33 \). The interaction is plotted in Figure 3.

**Response Time**

A similar analysis was conducted for item response time. As before, there was a statistically significant difference in response times between items, \( F(17, 4,675) = \)
and this has been explored more fully in the preceding sections. There was also a statistically significant main effect for group, $F(1, 275) = 665.91, \text{MSE} = 2,178.52, p < .001$, with the overall tendency being that university students took longer to respond than the school children. These effects are however qualified by a statistically significant interaction, $F(17, 4,675) = 17.01, \text{MSE} = 260.61, p < .001$. That is, the difference between the mean response time for university students and school students is not constant across the 18 items. Figure 4 plots an elaboration of this interaction between items and group membership. Items have been ordered along the abscissa according to their calibrated difficulty in the school sample.

Two aspects of Figure 4 are worth commenting on. First, there is a tendency for RTs to be more consistent for the school sample than the university sample, whose RTs tend to vary much more (Items 10, 11, and 14 have particularly long average response times and large standard deviations; see Table 3). However, the interaction between items and groups remained statistically significant with Items 10, 11, and 14 removed. Second, the difference between the two samples tends to increase as the calibrated difficulty of the items increase. The regression lines for each sample are represented in Figure 4, and there is a statistically significant divergence as calibrated difficulty increases, $t(32) = 2.19, p = .036$ (school: $b_{\text{School}} = 2.05, t_{[16]} = 3.35, p = .004$; university: $b_{\text{University}} = 7.15, t_{[16]} = 3.19, p = .06$. Note: $b$ = unstandardised regression coefficient).
These results indicate that when the 18-item LST is administered to a younger population with a larger age range, the influence of RC followed virtually an identical pattern of results to the more homogeneous sample of undergraduate university students. Although school students tended to perform poorer overall in terms of accuracy, they tended to respond more quickly. An analysis of the speed-accuracy trade-off effect (correlation between item RT and item accuracy) helps clarify this (Item 4 was omitted because all university students answered it correctly). The average speed-accuracy correlation across the 17 items was .16 (SE = .020) for the school students and .05 (SE = .036) for the university students. A pragmatic interpretation of this is that school students were inclined to favor a speedy response and make a guess on the more complex items after only partial reasoning.

General Discussion

The findings from both studies are comparable and indicate that increasing RC in the LST is associated with an increased number of errors and longer response times. Our findings suggest that the number of processing steps is also a significant predictor of both errors and response times, but only after the effect of RC has been considered. There was no interaction between RC and number of processing steps. Almost by definition, increasing the complexity of the reasoning in the LST is associated with the ambiguity generated by incomplete cells. We therefore partially expected that this might be a contributing factor in performance. In terms of errors, this was not the case. However, a greater number of incomplete cells did tend to increase the time to respond over and above that contributed by RC and number of processing steps. These results appear to generalize to a younger population of students. Compared to the university sample, the number of errors was greater in the younger age group, as expected. The school samples’ relatively quicker response times suggests guessing after partial reasoning on more complex items. In spite of this and its possible effect on reliability, these items were still psychometrically separable as indicated by good fit to the Rasch model and appropriate discrimination as indicated by the CTT point-biserial item-total correlations.

Alternative Theoretical Accounts

The processes that we have specified and modeled as features of the LST were theoretically driven by RC theory. The pattern of results provides convergent support for our analyses. Of course, it may be the case that alternative process theories might equally well account for the empirical data independent of our RC analyses. Here we consider a conceptualization that may not require the levels of relational integration we have proposed.

Consider the formatted ternary problem reported in Figure 1. We represent the processing as AND(B1(triangle), B4(circle), D2(square) → B2(cross)). A suggested solution protocol for this item is as follows: “There is a triangle and a circle in Column
B, so the target cell must be either a square or a cross. But there is a square in the second row so it cannot be a square. Therefore it must be a cross."

It might be argued that solution is possible by processing two or three (binary) steps in series where the output of one step is used as input for the next. The protocol itself is highly plausible and one that we believe would be used frequently. However, we are yet to be convinced that solution does not entail simultaneous processing. Our reasoning is as follows: The possibility of the target cell being either a triangle or a circle is eliminated by their presence in Column B. However, there is still ambiguity at this stage of the processing as a determinate solution is not possible—the target cell is either a square or a cross, but we do not know which. To solve the problem, the reasoner needs to consider sources of information beyond Column B, but not independent of what is already known about the elements in Column B—this is a requirement of the Latin square rule. With some searching of neighboring cells, the square in D2 will be discovered. If the rule is understood, it will be clear that this element must constrain the possible values in the target cell (as in fact any element in Row 2 will). Yet this new piece of information alone (i.e., not-square) cannot provide a determinate solution for the target cell. It needs to be considered in conjunction with the current representation of the task (i.e., square or cross) to derive a solution. Essentially what we are musing with here is the idea that parallel processing in practice does not mean that the representation of the task has to be built instantaneously (as the computational model of Halford et al. [1998a] does). Pertinent elements can be brought together over time to build a single representation—a relation. If it is only at this time that a solution is possible, then the structure of the relation at this point determines the complexity.

It is important to note that this is conceptually different from the idea of segmentation in multistep Latin square problems. In these problems, each intermediate step results in a determinate solution of one or more cells whose value is necessary for solution of the target cell (or other empty cells). If there were ambiguity in the value of the cell in an intermediate step then it would not qualify as a separate "segment"—the values of other cells would be needed to determine its status. This is consistent with the chunking and segmentation principles of MARC. Where variables interact, segmentation or chunking is not possible because information about the relationship between the variables needs to be processed.

Methodological Issues

For a new test to serve as a valid measure of anything, it is traditional that one first demonstrate reliability (Nunnally, 1978). A potential criticism with the current version of the LST is that the separate studies reported here do not yet demonstrate sufficient reliability of .80 or above. This is a direct result of trying to integrate two very different approaches to test construction. The first is based on a strong cognitive theory of how information is processed in working memory. The second is based on more atheoretical psychometric criteria. We believe that this integration, although difficult, is well worthwhile. Many psychometric indices are plagued by sample dependency. Reliance on any one indicator is problematic, and this is clear in the current study. The
reliability of the LST for one sample was .72 (18 items) and for the other .56 (17 items), which may give the impression that we do not yet have a sufficient basis for measurement. However, other information is necessary to place these statistics in an appropriate context. The modest reliability was based on a homogeneous sample of first-year university students all studying psychology. The higher reliability is from a school sample that is much more diverse in both age and presumably academic ability (the schools were not selective). When these samples were combined, the reliability approached the .80 rule of thumb (and would exceed .80 if more items were theoretically added).

It is clear that some items were too easy for the university sample and some too hard for the school sample. This serves to reduce reliability and is the inevitable cost of a test that is designed mainly on cognitive process and complexity theory. However, the potential utility of the LST beyond this is indicated in several ways. First, the analyses suggest that overall, the data provides a good fit to the Rasch measurement model in both samples—infit statistics indicate that individual item- and person-fit was good. Second, the rank order of the calibrated item difficulties across distinct samples was almost perfect, suggesting item calibration approached sample invariance. This would seem unlikely to occur if the only source of unreliability was random error. Third, because the development of LST items is theoretically driven and algorithmically derived, additional items at a given level of complexity and number of steps can be produced relatively easily to match the individuals being assessed. This has significant practical advantages for adaptive-testing situations.

Conclusion

The LST is theoretically derived and psychometrically applicable to a wide range of ages because it minimizes the role of knowledge. The complexity manipulation does not add rules to make the task more complex; instead it requires an increasingly complex instantiation of only a single rule. We have demonstrated robust a priori complexity effects in a collection of items that operationalize and differentiate parallel (relational) and serial processing in working memory. This is a new and very interesting task, and although the current investigation is still very introductory, we are optimistic that future research into this approach to assessing relational reasoning aspects of working memory will be fruitful. This study contributes to growing support for the RC theory as an account of working memory limitations and provides a good foundation for further exploration of relational reasoning and related constructs such as fluid intelligence.

References


