We propose that working memory and reasoning share related capacity limits. These limits are quantified in terms of the number of items that can be kept active in working memory, and the number of interrelationships between elements that can be kept active in reasoning. The latter defines the complexity of reasoning problems and the processing loads they impose. Principled procedures for measuring, controlling or limiting recoding and other strategies for reducing memory and processing loads have enabled complexity and yielded orderly quantification of capacity limits in both memory and reasoning. We argue that both types of limit might be based on the limited ability to form and preserve bindings between elements in memory.

Introduction

The techniques that humans have for making the best use of available information-processing capacity are of immense value but they have to be controlled to study capacity limitations and the effects of complexity. New, principled procedures for measuring, controlling or limiting recoding and other strategies by which subjects reduce memory and processing loads have enabled complexity and capacity effects to be investigated independently of, and in interaction with, knowledge. Here, we present the hypothesis that this development enables a new, integrated treatment of reasoning and working memory (WM), including an orderly quantification of capacity limits, and that this has opened up new research opportunities.

WM and reasoning

Developments in both theory and methodology have strengthened the links between WM and reasoning, and some salient points are summarized in Box 1.

We propose that the essential link between WM and reasoning is in the common requirement to bind elements to a coordinate system. Consider, first, short-term serial recall of the words ‘Fido, Rover, Cleo’. The words are assigned to ordinal positions when presented (Figure 1a) but this assignment must be maintained for later recall, and this requires attention. Even in free recall (not shown), items on a trial must be bound to the present trial concept or node in memory; binding might be even more extensive, inasmuch as an associative network between items would greatly facilitate recall. Now consider a choice reaction-time task, in which participants press a different button in response to one of several lights, and the buttons are assigned to lights randomly (noncompatible mappings; Figure 1b). Maintaining the binding of button positions to light positions in the WM requires attention [1]. Finally, consider a transitive inference problem, such as ‘Jane is taller than Wendy, Amelia is taller than Jane’. This can be solved by mapping premise elements into an ordering schema, as shown in Figure 1c.

Maintaining bindings between elements and slots using attention is common to WM and to reasoning. This is not the only common factor but there is substantial evidence that WM capacity (WMC) accounts for a sizeable proportion of the variance in reasoning [1,2] and intelligence [3]. WM and reasoning differ in whether the binding is supplied with the input (as in short-term serial recall) or has to be constructed by the reasoner, as in syllogistic (including transitive) inference, where premise elements have to be mapped to slots in a mental model in a way that is consistent with the premises. There are intermediate cases, such as where the mapping of responses to inputs, although predetermined, must be coded by the performer, or where participants construct their own mnemonics. Coordinate systems in the WM are less likely to include explicit relational representations than those in reasoning. Use of an explicit symbol that differentiates different kinds of links is a feature that distinguishes relations from associations [4]. In Figure 1c, the relation ‘taller than’ is explicit, whereas the ordering schemes used in the WM might not be.

We further propose that the common demand for attention when binding elements into slots is a possible explanation for common capacity limitations in WM and reasoning (Box 2). Here, rather than presenting a detailed model of these processes [4], we focus on methods for controlling chunking, recoding, rehearsal and other strategies, so that capacity limits can be measured.

WM capacity

Immediate memory capacity is roughly constant if measured in meaningful units, termed ‘chunks’ [5]. For example, FBICIAIRS comprises up to nine chunks, but only three chunks if the observer recognizes three familiar acronyms (FBI, CIA, IRS). WMC can only be estimated if the number of chunks can be determined, and techniques for this are summarized in Box 2.
A review [6] of disparate procedures in which items cannot be grouped or rehearsed (Box 2) converges on a range of 2–6 items in adults (usually 3–5), and fewer in children or the elderly[7]. In cases of exceptional or expert memory, it is apparently the size of recalled chunks, rather than their number, that is extraordinary [8].

Measuring complex chunks
A second approach to assessing capacity limits in chunks (Box 2) is to manipulate and estimate chunking processes

(a) Working memory

<table>
<thead>
<tr>
<th>Temporal position</th>
<th>First</th>
<th>Second</th>
<th>Third</th>
</tr>
</thead>
<tbody>
<tr>
<td>Items</td>
<td>Fido</td>
<td>Rover</td>
<td>Cleo</td>
</tr>
</tbody>
</table>

(b) Choice reaction time

Stimuli

Response buttons

(c) Transitive reasoning

Premises: Jane is taller than Wendy
Amelia is taller than Jane

Conclusion: Amelia is taller than Wendy

<table>
<thead>
<tr>
<th>Mental model</th>
<th>Tallest</th>
<th>Middle</th>
<th>Shortest</th>
</tr>
</thead>
<tbody>
<tr>
<td>Premise elements</td>
<td>Amelia</td>
<td>Jane</td>
<td>Wendy</td>
</tr>
</tbody>
</table>

Figure 1. Binding of elements into coordinate systems in (a) WM, (b) choice reaction time and (c) reasoning. (a) In WM, the coordinate system corresponds to spatial or temporal position. The identity of items in each location or position must be maintained in the presence of distraction or updated during the task. (b) The coordinate system corresponds to spatial location. In the non-compatible arbitrary condition, the left–right position of the stimuli is not consistent with the position of the response buttons. Temporary bindings between stimulus and response representations must be constructed and maintained in the presence of irrelevant information (colour). This imposes demands on attentional capacity. (c) Relational premises are integrated to generate an ordered triple. The coordinate system corresponds to the ordinal position. Each position functions as a variable.

Central capacity
Theoretically, capacity limits could occur separately in different parts of the processing system. Recent evidence begins to address questions such as (i) whether there is a central capacity that can be allocated across different modalities and codes, and (ii) where the capacity might reside in the brain.

One can combine a spatial memory task with a verbal task to determine whether the amount retained from one task affects the amount retained from the other. Sometimes,
Observed [12,13], consistent with the view that storage is little interference between verbal and spatial tasks is
across modalities. Thus, the central capacity could be allocated
to between modalities emerged. Subjects could retain either
spoken digits and coloured squares, and showed that, if
are (e.g. overt verbal retrieval). In unpublished research,
are not part of the store used for visual memory (e.g.
spoken aloud. This suggests that some mnemonic processes
are to be recalled from a meaningful category (e.g. states). Each burst of items recalled is assumed to
reflect WM filled to capacity from long-term memory once [48].
Memory updating. Items are to be remembered, some of which
are eligible to be changed (updated). Responses slow markedly as
eligible items increase from one to three [34].
Probed recall. Participants indicate if a probe occurred in a just-
presented list. At least three semantically related items near the
end of the list can show especially fast retrieval [49]. There is interference from other, similar lists if the number of items
exceeds basic capacity [50].
Multiobject tracking (examines attention limits related to WM). An
array of items includes several cues as targets. The cues
disappear; all items then move in different random trajectories. When they stop, the participant must identify the targets. Targets
up to the capacity limit can be tracked together as one changing
object [51].
Reconstruction by chunks. A structured configuration (chess
setup) is examined and reconstructed from memory. The pieces
reconstructed without pause presumably form a single chunk;
capacity is estimated as chunks recalled per look [52].
Recollection by chunks. Word series with varying interitem
association strengths are presented; words recalled in the
presented order are taken to reflect a single chunk [9,11].

Another study [16] examined memory for verbal-spatial
associations (names shown in schematic houses at different
screen locations). Adults could perform this task by
combining modality-specific memories for the name
sequence in the trial and for the spatial path between
the houses in which the names appeared, provided that
there was a one-to-one correspondence between names and
houses. When this strategy was precluded, crossmodal
associations presumably had to be retained. This occurred
when some houses contained two names and others con-
tained none, when verbal rehearsal of the names was
suppressed or when the subjects were children too young
to rehearse.

There is some evidence regarding where central
capacity might reside. The parietal lobes take part in an
attention system that integrates information across mo-
dalities, and might correspond to the seat of attention,
separate from frontal areas involved in controlling atten-
tion [17]. Although both areas are often active together in
WM tasks that involve the storage and manipulation of
information, transcranial magnetic stimulation (TMS) to
the frontal lobes disrupts only the manipulation, whereas
parietal stimulation disrupts storage as well [18]. Some
parietal areas show functional magnetic resonance imaging
responses to simple visual arrays that change across
memory loads in a way similar to behavioural responses
[19]. When feature complexity is added to the stimuli by
requiring memory for several features (e.g. colour in
addition to shape), some areas respond according to
complexity [20]. The brain data seem to be compatible with
a central store that reflects the focus of attention used as a
store [21], although it has not yet been proven that parietal
storage does reflect the focus of attention.

Chunk capacity and chunk size limits
People recall lists best if they are separated into groups of
2–4 items, the optimal typically being three [22,23]. It
could be viewed as coincidental that the optimal group
size is approximately the same as the maximal number of
chunks that can be recalled. However, it might be more
than coincidental, if the capacity limit is attention related.
Formation of a new chunk might be possible only if the
elements are retained in the capacity-limited store (focus of
attention?) concurrently, enabling rich associations to form
between the items [21]. The optimal group size is therefore
the largest number of items that can be reliably stored to be
interassociated at once. Chunks of any size can be built up,
of course, using the capacity-limited region reiteratively.
The fact that larger chunks can be built up with practice
and attention, and that the WM no longer seems to limit
performance [8], does not mean that chunk capacity limits
cease to have a role; expertise leads to larger chunks [24]
(liek the configurations in a chess game) but, when inde-
dependent estimates of capacity are obtained, they still equal
≈3–4 chunks in adults, and fewer in children [3,25].

Links between WM and reasoning
Accumulated evidence indicates that estimates of WMC
have some degree of domain generality and have a central
tendency of ≈3.5 items. We hypothesize that the WM limit
reflects capacity for attention, which determines the number of elements, whether items or chunks, that can be bound into a coordinate system. This hypothesis links WMC directly to a theory of capacity limitations in reasoning, relational complexity (RC) theory [4], considered next, which proposes that reasoning is limited by the number of variables or slots that can be related in a single representation.

**Complexity in reasoning**

Measures of the number of symbols required to define a concept [26,27] have achieved significant successes, as have metrics based on rule hierarchies [28]. Complexity metrics have also been important in cognitive development [29]. However, we focus on a metric that discriminates between complexity and knowledge effects, and which incorporates binding of problem elements to mental models.

**The RC metric**

Complexity can be defined by the number of slots or ‘arity’ of a relation that must be represented to perform a specific cognitive process (for example, a binary relation, such as ‘larger than’ has two slots, one each for the larger and smaller entities) [4,30–32]. Each slot of a relation can be filled in a variety of ways, and corresponds to a variable or dimension, and an n-ary relation is a set of points in n-dimensional space. RC norms are summarized in Table 1. The relation is a coordinate system, as noted earlier, so RC corresponds to the number of bindings to a coordinate system, consistent with our suggestion of how capacity is defined in the WM.

Conceptual chunking involves recoding concepts into less complex relations. For example, speed = distance/time, is a ternary relation but speed can be recoded into a unary relation, speed(60 kph), as when speed is indicated by a pointer on a dial. Although chunking reduces processing load, chunked variables cannot be accessed. Thus, the chunked representation of speed does not enable us to answer questions such as, ‘how does speed change if we cover the same distance in half the time?’, and we must revert to the ternary relation to calculate the answer. (Similarly, if one needs to access part of a chunk in the WM or its internal structure, the chunk must be unpacked.) Conceptual chunking is analogous to mnemonic chunking, in that it compresses a representation into fewer variables, but explicit relations are defined between the variables, and they are not simply associated (e.g. velocity is not an association between speed, distance and time but a specific relation defined between them).

‘Segmentation’ entails breaking tasks into less complex steps, which can be processed serially. Strategies and algorithms are common ways of doing this (e.g. adding one column at a time in multidigit addition). Representing complex structures hierarchically, and processing one level of the hierarchy at a time, is also an effective way to segment tasks [28]. Chunking and segmentation skills are important components of expertise.

There is common ground between WM and reasoning. First, both are limited by the ability to map elements into coordinate systems, which in turn depends on attention [33,34]. Second, both involve compressing material into chunks, mnemonic or (for reasoning) conceptual in nature. Conceptual chunks contain more relational information than do mnemonic chunks because the coordinate systems in reasoning are explicit relations. Third, the limitation is in the number of independent components, irrespective of their size. Chunk size in the WM is analogous to the number of possible instantiations of a relational slot in reasoning (i.e. whether a slot can be instantiated in 2, 3, …, or arbitrarily many ways). In both cases, this factor has less influence on capacity than does the number of chunks in the WM, or the number of slots in reasoning.

**Method for analysis of RC (MARC)**

MARC principles that have been found to be consistently valid across many domains are shown in Box 3. Application of these principles shows some tasks are difficult because they resist decomposition. Two examples are discussed here (Figures 1c,2).

**Transitive inference** imposes a processing load [30] because both premises need to be considered to assign premise elements uniquely into slots in the representation in the WM (Figure 1c). There are three slots, so RC = 3. Premise integration in transitive inference involves neural circuits in the prefrontal and parietal cortices [35,36]. Prefrontal regions are also activated in other reasoning

![Figure 2](https://www.sciencedirect.com)

**Figure 2.** An application of the MARC to the knight and knaves reasoning task.
Box 3. Principles of the MARC

Complexity analyses begin with established models of reasoning processes, taking account of knowledge and strategies likely to be available to the reasoner.

Assessments must be appropriate for the participant sample, and control tests are essential (e.g. closely matched binary-relational tasks are important controls for assessing ternary relations) [30].

1. Effective RC for a cognitive process is the least complex relation required to represent it. The required mental models are determined from established process models, supported by empirical research.

2. When tasks entail more than one step, the processing complexity of the task is the relation that must be represented to perform the most complex step, using the least demanding strategy available to humans for that task [4].

3. Variables cannot be chunked or segmented if relations between them must be processed because relations between chunked or segmented variables become inaccessible. This leads to several corollaries:
   - Chunking sets. The relation between A and both B and C is binary if the relation between B and C can be chunked, which is possible if the relation(s) between them need not be processed. In oddity tasks, a strategy of noting that red is different from three blues has RC = 2 [53].
   - Interaction between variables. Variables cannot be chunked or segmented if they interact because the influence of each variable is modified by the others. Therefore, interacting variables must be processed jointly.
   - When two arguments of a relation function as a unit for the cognitive process being performed, they can be chunked. In Figure 2 (of the main text), ‘A is a knave and B is a knave’ is one conceptual chunk for the first step. Similarly, if A is a knight and A says (some proposition), then the predicates ‘is a knight’ and ‘says’ can be chunked because they are a unit, equivalent to ‘The knight, A says’ [41].

4. When a cognitive overload cannot be reduced by conceptual chunking or segmentation, it is likely to be handled by default to a simpler representation, usually by ignoring one or more variables, leading to fragmentation of knowledge.

Box 4. Outstanding questions

Techniques for the quantification of reasoning complexity and the assessment of WMC imply potential advancements in several areas. Some of these have begun to be realized but there is scope for further research. The following questions outline some possibilities:

- **WMC.** Does the WM limit of 3–5 items reflect the focus of attention? Is there a pool of capacity that is general across domains? Is there also a limit on how long ideas remain active when no longer attended, or not [54]? If not, then what is special about the attended information? Is it that only information in the focus of attention can be bound together [6,55]?
- **WM and reasoning.** Are all possible interrelations between items concurrently in the focus of attention? Is that why WMC is limited [6,55]? Does such mutual binding enable new chunk formation and, simultaneously, comprehension?
- **Locus of capacity.** Where in the brain might central capacity reside? In the parietal cortex [18]?
- **Theoretical integration.** Complexity analyses are bringing a more orderly interpretation to cognitive developmental findings [56]. Does this resolve apparent conflicts, such as those between precocity and developmental change?
- **Classical findings.** Hirst et al. [57] demonstrated an improvement in concurrent tasks with practice, apparently indicating that there was no capacity limit, but this might not hold if expertise and WMC were assessed independently. Chi’s findings of superior retention by child chess experts, as compared with adult chess novices, demonstrates a powerful effect of knowledge on short-term memory [58]; however, when capacity is measured, it has been found to increase with age. Should these findings be reassessed in the light of contemporary knowledge of WM?
- **Assessment criteria in reasoning.** Given that the norms of logic are no longer seen as appropriate criteria for human reasoning, does cognitive complexity provide an alternative basis for assessment [59]?
- **Transition processes.** Increasingly complex concepts can be acquired by adding variables to representations (e.g. density helps to account for variations in size and weight). Acquisition depends both on knowledge and on capacity to process the extra dimension; however, chunking and segmentation can reduce processing loads. Do these factors have implications for transition processes in cognitive development?
- **Improving experimental design.** The effects of complexity are not always recognized. For example, the difference between categorization by rules and by information integration could be accounted for by complexity differences [60]. Should the effect of complexity be considered in experimental design, to avoid confounding by other variables?
- **Behavioural and brain imaging research.** Recent research suggests that brain regions such as the rostral lateral prefrontal cortex are more selective to cognitive complexity than to cognitive domain [39]. Can precise manipulation of cognitive complexity, with other factors controlled, increase the precision of brain imaging research?

Summary

We have proposed similar limitations in WM and reasoning, which seem to reflect a central capacity limit. WM is limited to ~3–4 chunks, and reasoning is limited to relations between four variables. One priority is to identify the common underlying mechanisms. A promising hypothesis is that both WM and reasoning require items or concepts to be the focus of attention concurrently. This would enable them to be interassociated to form chunks or bound into coordinate systems, including relational representations that enable inferences to be made. Another priority is to determine the underlying brain systems. Frontoparietal networks have been implicated. Parietal regions seem to be most important for focus of attention and storage and fluid intelligence (Gf) tasks that require relational integration, including modified Ravens Matrices problems [36,37] and analogy problems [38]. Moreover, as task complexity increases, more anterior regions seem to be recruited [39,40].

**Knight and knave** problems are segmented into separate inference steps (Figure 2) but each inference entails compound propositions. In step 1, the proposition knight(A) contradicts the proposition knave(A), which implies that knight(A) is false. Applications of the MARC [41] result in three conceptual chunks: ‘Knight A says’, ‘A and B are knaves’ (as a chunked argument of ‘says’), and ‘the hypothesis is false’. Therefore, for this step RC = 3.

The need to interpret inputs jointly constrains decomposition into simpler subtasks. Some tasks that are difficult or complex (e.g. according to cognitive complexity and control theory [28]) incorporate this factor [42]. Interpretation of interactions was used to determine how many variables humans can process in parallel, while controlling techniques for reducing processing load. Performance declined as complexity increased, and was at chance level on five-way interactions [43]. Given that an n-way interaction is defined on n variables, so content is assigned to n slots, the data indicate that humans are limited to processing four variables.
functions of WM, whereas prefrontal regions seem to be more important for manipulation functions in WM and reasoning. If manipulation increases RC, then TMS applied to prefrontal regions should be more disruptive as the RC of a task increases. Procedures for controlling simplifying strategies have opened the way for these and other lines of investigation into WM, reasoning, the links between them, and the underlying brain systems (Box 4).

References
5 Miller, G.A. (1996) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol. Rev. 63, 81–97
23 Wickelgren, W.A. (1964) Size of rehearsal group and short-term memory. J. Exp. Psychol. 68, 413–419
43 Halford, G.S. et al. (2005) How many variables can humans process? Psychol. Sci. 16, 70–76
55 Davis, G. et al. (2001) Can attention select only a fixed number of objects at a time? Perception 30, 1227–1248
Perceptual Development (Kuhn, D. and Siegler, R., eds), pp. 557–608, Wiley


Have you contributed to an Elsevier publication?

Did you know that you are entitled to a 30% discount on books?

A 30% discount is available to all Elsevier book and journal contributors when ordering books or stand-alone CD-ROMs directly from us.

To take advantage of your discount:

1. Choose your book(s) from www.elsevier.com or www.books.elsevier.com

2. Place your order

   Americas:
   Phone: +1 800 782 4927 for US customers
   Phone: +1 800 460 3110 for Canada, South and Central America customers
   Fax: +1 314 453 4898
   author.contributor@elsevier.com

   All other countries:
   Phone: +44 (0)1865 474 010
   Fax: +44 (0)1865 474 011
   directorders@elsevier.com

   You’ll need to provide the name of the Elsevier book or journal to which you have contributed. Shipping is free on prepaid orders within the US.

   If you are faxing your order, please enclose a copy of this page.

3. Make your payment

   This discount is only available on prepaid orders. Please note that this offer does not apply to multi-volume reference works or Elsevier Health Sciences products.

   For more information, visit www.books.elsevier.com